Родригес Залепинос Р.А. Данные и методы интеллектуального анализа данных для исследования окружающей природной среды. Системный анализ и информационные технологии в науках о природе и обществе (САИТ-2011), стр. 94–107. Выпуск 1 – Донецк: ДонНТУ, – 2011. – 214 с.
Электронная версия статьи: http://www.csm.donntu.edu.ua/rodriges/publications/rodriges_env_data_mining_2011.pdf

УДК 004.853

Данные и методы интеллектуального анализа данных для исследования окружающей природной среды

Родригес Залепинос Р.А.
Донецкий национальный технический университет,
rodriges@csm.donntu.edu.ua

Abstract

Rodriges Zalipynis R.A. Data and data mining methods for natural environment research. This paper gives a comprehensive survey of the state-of-the-art data mining techniques developed to help Earth scientists to advance in their efforts to better understand the environment. The paper also tries to bridge the interdisciplinary gap between computer and Earth scientists by reasoning the application of data mining techniques and explaining mechanisms of the underlying environmental phenomena. First, author analyzes available environmental data along with the ways of collecting, storing and processing it, including Ukrainian network for collection of meteorological data, ground, marine and satellite data sources, vegetation indices and reanalysis archives. Other sections are devoted to the discovery of teleconnections with shared nearest neighbor clustering algorithm and deriving association rules between anomalous events abstracted from meteorological parameters and vegetation indices with Apriori algorithm. The paper can serve as a good place to start from learning about this interdisciplinary research.

Введение

Данная статья является первым обзором данных и методов интеллектуального анализа данных для исследования окружающей природной среды. Во-вторых, для исследуемых климатических феноменов и рассмотренных методов, приводятся их описания и обоснование применения, синтезированные из многих источников по климатологии и метеорологии, на которые не ссылались оригинальные статьи. С точки зрения автора, это поможет сократить междисциплинарный разрыв между исследователями в области компьютерных наук и наук о Земле.

Тот факт, что погода в одном районе может быть связана с погодой где-то ещё, довольно далеко от данного места, всегда привлекал внимание людей. Многие экстремальные явления можно предугадать, анализируя такие дальние корреляционные связи и закономерности их появления. Это позволяет вовремя принять соответствующие меры по эвакуации людей, спасая их жизнь и имущество при возникновении чрезвычайных ситуаций (ураганы, наводнения, оползни и т.п.).

В разделе «Данные» проанализирована организация глобальной системы наблюдений всемирной службы погоды, географическое распределение метеостанций, система сбора метеоданных в Украине, спутниковые методы сбора данных, на основе которых можно вычислять индексы растительности, и архивы повторного анализа с открытым доступом.

Методы получения таких связей и описания закономерностей их появления рассмотрены в разделе «Поиск телеконнекций». Телеконнекции существенно влияют на местную погоду, часто вызывая засухи, ливни, наводнения, периоды сильной жары и холода, что наносит вред сельскому хозяйству, водоснабжению и рыболовству. Более того, они могут влиять на качество воздуха, пожароопасность, энергоснабжение и здоровье людей [6, стр. 286].

Знание о существовании телеконнекций, их характере, изучение их природы и изменений в их поведении служат ключом в понимании изменений и изменчивости регионального климата.

Климат является сложной системой, компоненты которой взаимодействуют друг с другом иногда неожиданным образом. В разделе «Ассоциативный анализ» приводится подход, позволяющий получить закономерности между аномальными значениями метеорологических величин (температуры, осадков и солнечной радиации) и индексов растительности, рассчитанных на основе спутниковых данных.

Разделы «Поиск телеконнекций» и «Ассоциативный анализ» независимы, однако для понимания обоих обязателен раздел «Данные».

Проблематика

Представьте себе шар радиусом более 6000 км, с желобами до 11 км и пиками до 9 км, площадью 510×10^6 км^2, 3/4 поверхности которого занимает среда, в глубинах которой побывало меньше человек, чем в космосе, заполненная наименее химически изученным веществом на планете. Оставшаяся территория крайне не-

Rodriges Zalipynis R.A. Data and data mining methods for natural environment research. System analysis and information technology in environmental and social sciences (2011), pp. 94–107. Donetsk National Technical University. Volume 1. Donetsk, 2011. – 214 p.

равномерно устлана почти 300 000 видами растений, льдом, асфальтом, песком и населена более 1.2×10^6 видами животных. Представьте также, что шар окутывает слой толщиной 3000 км и массой $5{,}27 \times 10^5$ тонн разных газов, которые вместе вращаются вокруг оси шара со скоростью более 1500 км/час. Все это неравномерно нагревается ядерным реактором (Солнце) на расстоянии 150×10^6 км.

Задача исследования климата состоит в том, чтобы как можно больше узнать обо всех описанных компонентах и процессах, их взаимодействии, особенностях и закономерностях, понять все это и описать математическими зависимостями.

Задача прогноза климата и погоды состоит в том, чтобы используя полученные закономерности и всю доступную на текущий момент информацию предсказать состояние атмосферы на месяц, сезон, год и 100 лет вперед.

Погода – это совокупность значений метеорологических параметров и явлений, которые характеризуют состояние атмосферы в конкретном регионе планеты в определенный момент времени [1].

Климат – «средняя» погода: средние значения метеорологических величин и степень их изменчивости за определенный период времени (обычно от месяца до миллионов лет) [6, стр. 96].

Интеллектуальный анализ данных (data mining) – «статистика в масштабе, скорости и простоте» [2, слайд 4] – процесс автоматического извлечения интересных, скрытых, неявных и потенциально полезных закономерностей из больших объемов данных [3]. Если данные описывают процессы и объекты, имеющие географические и временные размеры, то для них разрабатывают методы пространственно-временного интеллектуального анализа данных [52, стр. 240].

Современный мир испытывает взрывообразный рост количества данных, генерируемых быстрее, чем их успевают проанализировать. Науки о Земле не являются исключением.

Более чем за сто лет (1901–2008 гг.) Национальный климатический центр данных США накопил 1.7×10^9 измерений от 30×10^6 приборов [4, слайд 21].

Спутники НАСА наблюдения за землей (Earth Observing Satellites, EOS) собирают около 1-го терабайта данных ежедневно [52, стр. 237].

Несколько терабайт данных – типичный выход модели прогноза климата на 100 лет с 6-ти часовым шагом для более 100 переменных глобальной решетки $1.4° \times 1.4°$. Для получения правдоподобного прогноза необходимо несколько сотен прогонов модели [5, стр. 2].

Ценность имеющейся информации на порядок ниже без эффективных методов ее анализа. Обширные базы климатических данных предоставляют беспрецедентные возможности для поиска интересных и полезных закономерностей, однако в этой области традиционный ручной подход «гипотеза–проверка» существенно ограничен в силу своей трудоемкости.

Методы интеллектуального анализа данных предоставляют автоматические, но при этом осмысленные подходы к генерации гипотез и их проверке, а также эффективные средства работы с массивами данных, которые не помещаются в оперативную память вычислительных систем.

Применение методов интеллектуального анализа данных позволяет получить известные закономерности, подтверждая тем самым адекватность разработанных подходов, а также ранее неизвестные закономерности, которые могут быть потенциально новыми природными феноменами.

Данные

В этом разделе анализируются доступные на сегодняшний день данные для изучения окружающей природной среды, способы их сбора и обработки. В литературе можно найти историю развития метеонаблюдений в России [7, стр. 6], синоптической метеорологии 19–20 вв. [8, стр. 27–35] и в античное время [9].

В 1967 г. Всемирный метеорологический конгресс принял план Всемирной службы погоды (ВСП). С тех пор и по нынешнее время ВСП состоит из трех основных компонентов [8, стр. 53; 10, стр. 8], являясь частью Глобальной системы систем наблюдения за землей (ГЕОСС) [11]: глобальной системы телесвязи (ГСТ), глобальной системы обработки данных (ГСОД) и глобальной системы наблюдений (ГСН).

Глобальная система наблюдений ВСП. Стационарные наземные метеостанции проводят наблюдения за атмосферным давлением, температурой и влажностью воздуха, скоростью и направлением ветра, формой облаков, количеством осадков, атмосферными явлениями (туман, гололед и т.п.). Эти же параметры измеряются на различных высотах наземными аэрологическими станциями вертикальным зондированием атмосферы с помощью радиозондов.

Автоматизированные метеостанции – компактные устройства для измерения метеопараметров в автоматическом режиме через заданные промежутки времени. Они могут передавать измерения посредством подключения к ЭВМ либо удаленными средствами связи.

На кафедре компьютерных систем мониторинга Донецкого национального технического университета установлена автоматическая метеостанция (см. рис. 1), измеряющая температуру, относительную влажность, атмосферное давление, направление и скорость ветра, коли-

чество осадков каждые 10 минут. Имеется также автоматический газоанализатор, измеряющий концентрации веществ NO_2, CO_2, SO_2. Графики доступны в системе ОМОС (Областная система мониторинга окружающей среды в Донецкой области) [12]. Станция не входит в ГСН ВСП.

Рисунок 1. – Автоматическая метеостанция кафедры КСМ ДонНТУ

По данным Национального климатического центра данных США (NCDC) [13] на территории современной Украины за 73 года (1936–2009 гг.) функционировало всего 202 метеостанции. Из них сегодня в действии только 47 со средней историей измерений в 60 лет (не считая 9-ти, которые работают с 2004 г.).

Для сравнения можно привести погодную службу Испании, у которой доступны данные по нескольким параметрам почвы, осадкам, максимальной и минимальной дневной температуре по 6750-ти станциям Испании и Португалии (Иберийский полуостров) и Балеарских островов с 1970 г. по настоящее время. Станции равномерно распределены по территории на расстоянии от 63 метров до 1 км друг от друга (2005 г.). В Польше таких станций 61, функционирующих с 1999 г. [14, стр. 220]. Не все станции входят в ГСН ВСП. Для сравнения: площадь Иберийского полуострова вместе с Балеарскими островами соизмерима с площадью Украины в отношении 0.997, а Польша почти в 2 раза меньше Украины [15].

В Украине наземная метеостанция представляет собой помещение и площадку с оборудованием для измерения метеопоказателей, а также персональным компьютером с программным продуктом «АРМ Метеоролога» [16], разработанным Харьковской компанией АО «Специальные системы связи».

Станция обслуживается человеком-оператором, который регулярно снимает метеопоказания с приборов на площадке наблюдения.

АРМ Метеоролога разработан с использованием языка программирования Delphi. Это графическое Windows приложение, которое облегчает ввод метеоинформации и передачи данных на региональный сервер. Например, на формах размещены списки множественного выбора для ввода категориальных признаков, обеспечивается контроль числовых параметров. Получаемые от оператора значения сохраняются в локальную базу данных под управлением MySQL. При отправке метеоданные кодируются в метеотелеграмму – строку цифр, букв и некоторых других символов для минимизации трафика.

Каждые 6 часов все наземные метеостанции мира одновременно в регламентированное международными соглашениями фиксированное время по Гринвичу передают свои измерения через Глобальную систему телесвязи (ГСТ) ВСП, которая имеет три уровня [10, стр. 11]. В итоге измерения от всех станций, входящих в сеть ВСП, поднимаясь от уровня к уровню, концентрируются в трех мировых центрах данных.

Со дня основания ВСП были учреждены три мировых метеорологических центра данных: в Москве, Вашингтоне и Мельбурне.

Национальный климатический центр данных США (National Climatic Data Center, NCDC) является одним из мировых центров данных по метеорологии. Он хранит архивы приземных, спутниковых и морских климатических данных приборов, входящих в ГСН ВСП [13]. В частности, через Интернет доступны измерения более 20 000 метеорологических станций (см. рис. 2) с 1901 г. по настоящее время с шагом до 1-го часа.

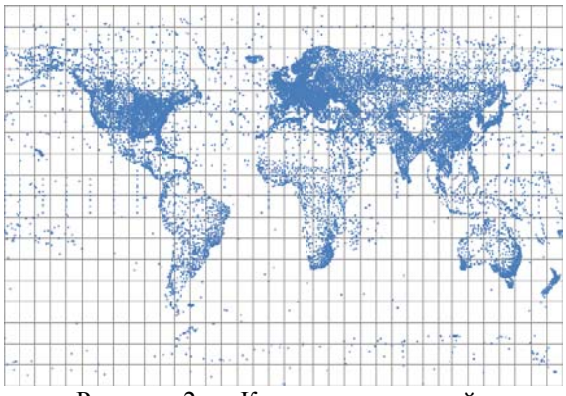

Рисунок 2. – Карта метеостанций, измерения которых находятся в базе данных NCDC [4, слайд 21]

Уровень национальной сети метеорологической телесвязи функционирует в пределах каждой страны. Например, метеостанции Донецкой области передают свои измерения в Донецкий гидрометеоцентр. Также в Донецк поступают данные из Луганского гидрометеоцентра. Данные из Донецка поступают в Харьков, а оттуда в Киев [17].

Уровень региональных сетей метеороло-

гической телесвязи объединяет национальные сети, а уровень главной сети телесвязи представляет собой высокоскоростную сеть, которая связывает три мировых центра данных и ряд узлов региональных сетей.

Связь в пределах Украинской национальной метеосети осуществляется по арендованным аналоговым телефонным линиям в режиме точка-точка. Для областных метеостанций также реализована возможность передачи данных на региональный сервер по GSM связи посредством мобильного телефона (покупается у мобильных операторов на общих условиях). Для передачи метеоданных по каналам связи Украинской национальной метеосети был разработан специальный протокол.

На областном сервере установлен программный комплекс «Бриз» [18], работающий под управлением Linux системы CentOS. «Бриз» осуществляет прием метеотелеграмм от областных метеостанций и передает их на следующий уровень национальной метеосети. При этом выполняется проверка метеотелеграмм на искажения при передаче, привлекая оператора, в случае необходимости. Также «Бриз» выполняет мониторинг состояний линий связи и участвует в приеме данных для «АРМ Синоптика».

Российская Федерация получила финансирование от Международного банка реконструкции и развития (МБРР) в сумме, эквивалентной 80 млн. долларов США [19] на техническое перевооружение всей наземной метеорологической наблюдательной сети Росгидромета.

По результатам конкурса исполнителем проекта является компания ЛАНИТ. Конкурс проводился Фондом "Бюро экономического анализа".

Проектом предусмотрены поставка оборудования для наблюдательной сети, состоящей из более 1900 объектов, и проведение работ на 240 площадках по всей территории РФ. ЛАНИТ обязуется поставить 210 автоматических станций и более 1600 автоматизированных метеорологических комплектов. Для измерения солнечной радиации будет установлена одна актинометрическая станция. В рамках проекта будет проведена автоматизация 85 региональных метеорологических центров, модернизация более 40 центральных и кустовых радирующих станций. Будет установлен широкий спектр современных средств связи: телефонные модемы, GPRS-модемы, устройства широкополосного доступа, низкоорбитальные спутниковые модемы, VSAT-терминалы, радиомодемы и др. на смену обыкновенной телеграфной и междугородней телефонной связи и каналам радиосвязи

Подобные проекты в таком масштабе в России ранее никогда не проводились [20].

Сбор метеоданных об океане. Мировой океан, занимая 3/4 поверхности планеты, играет ключевую роль в атмосферной циркуляции. Считается, что при долгосрочном прогнозе климата[1], взаимодействие между атмосферой и океаном служит наибольшим источником точности прогноза (skill) [35, стр. 498], поскольку океан – главный источник влаги, поступающей в атмосферу и огромный тепловой резервуар [8, стр. 42].

В океане метеоданные получают с островных гидрометеорологических станций (в основном не отличаются наземных станций), судов погоды (оснащены необходимой аппаратурой и средствами связи) [8, стр. 42], судов добровольного наблюдения и разного вида буев [21, стр. 147].

Дрейфующие буи измеряют температуру и горизонтальную скорость поверхности океана. Они следуют за поверхностными течениями и передают данные своих измерений через спутниковую систему АРГОС.

Заякоренные буи измеряют температуру, соленость, горизонтальное течение и биологические параметры на различных глубинах. Вертикальные профили температуры и солености измеряются обрывным батитермографом во время океанографических рейсов.

Вертикально ныряющие буи Арго измеряют профили температуры и солености, по мере того, как они опускаются и поднимаются в океане. Данные передаются через АРГОС во время нахождения буя на поверхности.

Хотя и реализуются многочисленные программы по контактному наблюдению за океаном, все же горстка существующих буев не позволяет выполнить глобальный комплексный мониторинг океана.

Спутниковые наблюдения предоставляют информации об океане и суше, собрать которую с поверхности планеты невозможно. Эти системы обладают широким обзором территории.

Космический компонент ГСН ВСП состоит из трех типов спутников: метеорологические низкоорбитальные, геостационарные и экспериментальные [10, стр. 9]. Спутники ТОПЕКС/Посейдон, Ясон-1, ERC-2, ЕНВИСАТ способны измерять цвет океана (концентрацию зоопланктона на поверхности) и уровень моря.

Большой интерес для исследования окружающей природной среды представляют спутники Национального управления США по исследованию океанов и атмосферы (National Oceanographic and Atmospheric Administration, NOAA). Спутник NOAA имеет высоту орбиты 870 км, совершает один виток за 102 мин., при котором удается получить информацию с поверхности около 3 000 × 7 000 км [22, стр. 120–

[1] Вероятностный прогноз на месяц с периодом упреждения 1⁄2 месяца и на 3 месяца с периодами упреждения от 1⁄2 месяца до 12.5 месяцев с шагом в 1 месяц [35, стр. 496].

122].

На борту спутника установлен усовершенствованный радиометр с очень высоким разрешением, AVHRR (Advanced Very High Resolution Radiometer).

Прибор способен принимать сигнал в окне прозрачности атмосферы 10-12 мкм, что позволяет оценивать температуру поверхности океана. Он также принимает сигнал в видимой и ближней инфракрасной областях спектра, что позволяет рассчитывать индексы растительности.

При этом прибор имеет разрешение 1.1 × 1.1 км и ширину полосы обзора 2800 км. Он способен составить полное изображение земли за одни сутки [22, стр. 94–96].

NDVI (Normalized Difference Vegetation Index) – нормализованный относительный индекс растительности, показатель количества фотосинтетически активной биомассы (обычно называемый вегетационным индексом). Данный индекс используется для решения задач оценки растительного покрова [23]. Вычисляется по формуле NDVI = (NIR – RED)/(NIR + RED), где NIR и RED – отражения в ближней инфракрасной и красной области спектра соответственно.

Расчет основан на том, что в красной области спектра (0,6-0,7 мкм) лежит максимум поглощения солнечной радиации хлорофиллом высших сосудистых растений, а в инфракрасной области (0,7-1,0 мкм) находиться область максимального отражения клеточных структур листа.

Области со здоровой либо густой растительностью поглощают большую часть видимого света и отражают большую долю инфракрасного излучения, а нездоровая либо разреженная растительность отражает больше видимого света и меньше инфракрасного. NDVI позволяет различать и анализировать виды растительности [24], см. рис. 3.

Первичная нетто-продукция (Net Primary Production, NPP) выражает количество CO_2, которое поглощается из атмосферы и перерабатывается растениями [25]. Временные ряды NPP для территорий отражают изменения в землепользовании (например, застройку территории), реакцию растительности на изменчивость климата (в том числе пожары и засухи) и изменения в объёме биомассы, которая является индикатором пищевой безопасности региона [26].

Потенциальная эвапотранспирация (Potential Evapotranspiration, PET) – количество воды, которое может быть испарено растениями при ее достаточном наличии. Измеряя PET можно неявно оценить показатели, от которых она зависит: тип почвы, глубины влаги в ней и тип растительности [27].

Спутниковые данные позволяют NASA реализовывать общественно полезные проекты. Например, анализатор условий произрастания сельскохозяйственных культур по всему миру (Crop Explorer), глобальная модель и система прогнозирования оползней, система предупреждения о пожарах, мониторинг наличия воды вдоль маршрутов перемещения скота для обеспечения сохранности пастбищ и другие проекты [28, стр. 26].

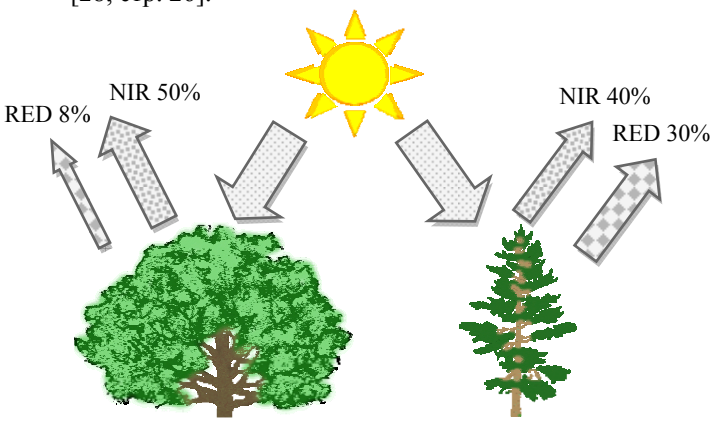

Рисунок 3. – Степень поглощения и отражения видимого красного (RED) и инфракрасного (NIR) излучений разными видами растительности

Проект Crop Explorer предоставляет открытый доступ к данным по атмосферным осадкам по всему миру, в том числе и по Украине. Данные можно визуализировать в системе Google Планета Земля в трехмерном режиме [29].

Архивы повторного анализа (reanalysis archives). Данные за историю метеонаблюдений получены разными приборами и способами, имеют разный характер (например, температура почвы в точке и температура водяного пара облаков над территорией), неравномерно распределены во времени и в пространстве и хранятся в разных форматах. Кроме того, в данных могут присутствовать неточности, погрешности и пробелы.

Весь набор информации трудно использовать, не проведя ее систематизацию.

Концепция архива повторного анализа возникла из потребности предоставить сообществу исследователей климата целостный ретроспективный ряд метеорологических состояний планеты, восстановленных на основе всех собранных данных за десятки лет метеонаблюдений.

В 1996 г. появился первый в своем роде архив повторного анализа NCEP/NCAR Reanalysis версии 1 (R1) [30, 31], созданный совместными усилиями Национальных Центров США по Предсказанию Окружающей Среды и Исследованию Атмосферы (National Center for Environmental Prediction, NCEP and National Center for Atmospheric Research, NCAR).

Были восстановлены поля атмосферных

переменных на регулярной широтно-долготной многоуровневой сетке, охватывающей всю планету при помощи ассимиляция данных, собранных с приборов на суше, кораблях, вертикального зондирования атмосферы (с помощью водородных баллонов), самолетах, спутниках и других данных (см. рис. 4).

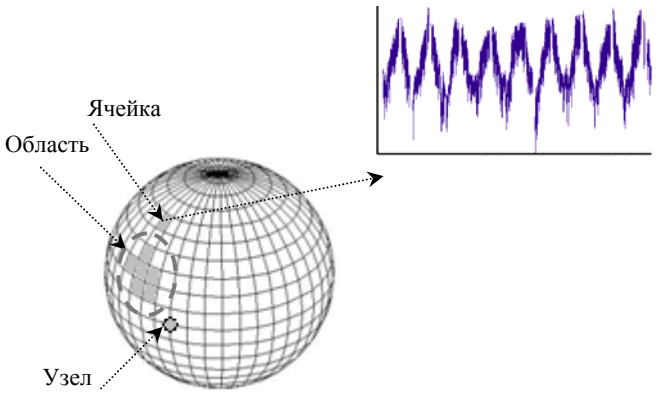

Рисунок 4. – Упрощенная иллюстрация представления данных на широтно-долготной решетке

Усвоение (ассимиляция) данных (data assimilation) – набор статистических и численных методов комбинирования всех доступных метеоданных для их интерполяции на регулярную широтно-долготную решетку.

Цель усвоения данных – как можно точнее определить состояние атмосферы в заданный момент времени [32, стр. 136].

Полный объем архива составляет 2.9 Терабайта (по состоянию на август 2009 г.). Архив содержит более 80 различных переменных (включая атмосферное давление на уровне моря, влажность воздуха, солнечную радиацию) в нескольких координатных системах с 1948 г. по нынешнее время [31].

Вторая версия архива под названием NCEP–DOE AMIP R2 отличается исправленными ошибками и использованием новых систем ассимиляции данных [33]. При этом архив содержит данные для более короткого срока – с 1979 г.

Поиск телеконнекций

В метеорологии телеконнекция (teleconnection, дальняя корреляционная связь) – существенное одновременное колебание климата в разных довольно далеко расположенных друг от друга географических районах [34, стр. 38].

Знания о телеконнекциях используются Центром Прогноза Климата (Climate Prediction Center, CPC) для сезонного и долгосрочного прогноза климата [35, стр. 497], изучения и прогноза загрязнений атмосферного воздуха [36], построения систем оповещения о приближении суровых погодных условий [37].

Самой знаменитой телеконнекцией считается Эль-Ниньо (El Niño, с исп. «младенец Христос») – появление аномально теплой воды на поверхности океана у берегов Перу в декабре или на Рождество (отсюда и название) [38, стр. 153]. Точность сезонного прогноза климата в большой степени зависит от точности прогноза наступления и степени Эль-Ниньо [39]. Большинство телеконнекций связано с океаном, поэтому он вызывает основной интерес при их поиске.

Механизм происхождения Эль-Ниньо таков. Мировой океан опоясывают особые северные и западные ветры (см. рис. 5).

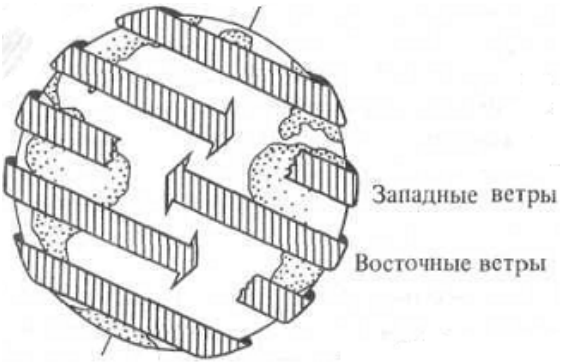

Рисунок 5. – Пояса ветров над Мировым океаном [38, стр. 29]

Восточные ветры, дующие примерно над экватором называются пассатами. В обычное время они перемещают теплую воду вдоль экватора к Азиатскому континенту, где уровень моря повышается примерно на 1 м по отношению к Американскому побережью [38, стр. 152], а на место ушедшей теплой воде из глубин поднимается холодная вода (см. рис. 6).

Рисунок 6. – Повышение уровня моря под действием пассатов [38, стр. 152]

Бывают периоды, когда происходит резкое ослабление пассатов. Большая возвышенность теплой воды у Азиатских берегов более не может удерживаться такими слабыми ветрами и устремляется назад. Для Индонезии, Австралии и юго-восточной Африки значительные явления Эль-Ниньо вызывают суровые засухи и опустошительные лесные пожары. Для Эквадора, Перу

и Калифорнии оно приводит к обильным ливням и ураганам, которые часто вызывают наводнения и оползни. Сильные всплески Эль-Ниньо приносят тысячи смертей, разрушают дома людей и причиняют миллиарды долларов убытка мировой экономике [40]; конкретный пример в [38, стр. 153] для Эль-Ниньо 1972 г.

Программный продукт NASA World Wind [41] способен в анимированном трехмерном режиме визуализировать аномалии температуры океана, полученные со спутника, во время Эль-Ниньо.

Сам Эль-Ниньо является следствием ещё более крупномасштабного Южного колебания (Southern Oscillation) – тенденции повышения атмосферного давления над Тихим океаном, при его падении над Индийским океаном [42] (что и вызывает ослабление пассатов). В свою очередь, Южное колебание вызывается процессами планетарного масштаба: взаимным расположением планет, солнца и луны [43, стр. 679–682].

Для наглядного представления поведения известных колебаний и вызываемых ими телеконнекций пользуются простыми в построении климатическими индексами океана – временными рядами климатических переменных, измеренных для участков океана, в которых периодически проявляются аномалии климата [6, стр. 287–295].

Индексы, построенные на основе температуры поверхности океана либо давления на уровне моря, называют климатическими индексами океана (Ocean Climate Indices, OCI).

Например, для Южного колебания традиционно используется разница атмосферного давления на уровне моря между Дарвином (Австралия) и Таити (см. рис. 7); отрицательные значения индекса соответствуют появлению феномена Эль-Ниньо.

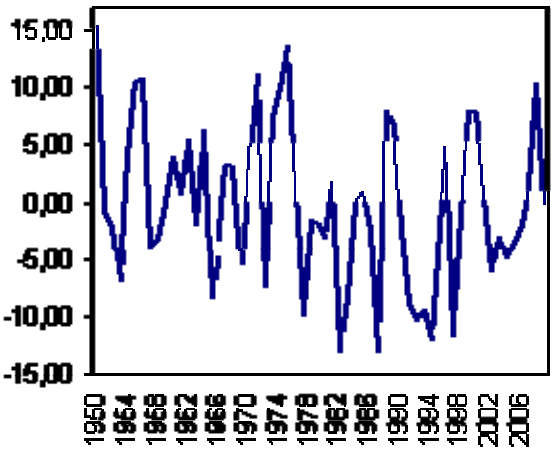

Рисунок 7. – Индекс Южного колебания (по данным Метеорологического Бюро Австралийского Национального Климатического Центра [44])

Климатическая система может содержать потенциально неизвестные ученым телеконнекции, для поиска которых климатологи использовали эмпирические ортогональные функции (Empirical Orthogonal Functions, EOF) [34, стр. 44–47].

В работах [45, 46] выполнен поиск телеконнекций с помощью кластеризации. Подход мотивируется тем, что возникающий климатический феномен захватывает существенную область океана либо суши, величина атмосферного давления либо температуры в которой однородна.

Кластеризация разбивает множество объектов на группы (кластеры), таким образом, чтобы объекты в одном кластере были более близки друг к другу, чем к объектам из других кластеров [47, стр. 490].

Значения временных рядов климатической переменной для ячеек широтно-долготной решетки представляются в виде векторов и подаются на вход алгоритму кластеризации SNN. Хотя вектора не содержат информации о координатах своих ячеек, в один кластер попадают соседние ячейки, формируя непрерывные регионы благодаря пространственной автокорреляции.

Алгоритм кластеризации SNN (Shared Nearest Neighbor) представляет вектора в виде вершин графа и определяет меру близости между двумя вершинами на основе количества их общих ближайших соседей. Плотность региона вокруг вершины оценивается её суммарной близостью к своим ближайшим соседям. На основе этого алгоритм сохраняет ребра в однородных по плотности районах и удаляет ребра в переходных зонах. Для одного кластера определяются несколько представительных вершин (имеющие наибольшую связность) [48, стр. 3–4].

Перечисленные особенности позволяют находить кластеры различных плотностей (как регионы с однородной плотностью), форм и размеров. Алгоритм также не кластеризует все вершины, классифицируя некоторые как шумы и выбросы, и автоматически определяет «естественное» количество кластеров в зависимости от данных и заданных параметров.

Алгоритм использует пороги $k, \Theta_1, \Theta_2, \Theta_3$ (подробнее шаги 2, 5, 6, 7 соответственно) и работает следующим образом [48, 49, 50].

1. Построение графа подобия. Вычислить расстояние между всеми парами входных векторов и построить полный граф, в котором вершины представляют собой вектора, а веса рёбер – расстояние между ними.

Пусть $Adj[i]$ – список смежности вершины i, $|Adj[i]|$ – количество элементов в списке $Adj[i]$, $Adj[i][j]$ – индекс вершины на позиции j в списке смежности вершины i, $w(i,j)$ – вес ребра (i,j).

2. Разрежение графа подобия. Для каждой вершины i упорядочить список $Adj[i]$ по возрастанию весов ребер: $w(i, Adj[i][j]) \leq w(i, Adj[i][j+1])$, $j = 0..|Adj[i]| - 1$. Удалить хвост списка $Adj[i]$, оставив первые k элементов.

3. Построение графа Общих Ближайших Соседей. Удалить вершину i, из списка $Adj[j]$, если $j \notin Adj[i]$ (такие случаи довольно часты после шага 2, особенно на границах соседних кластеров с разной плотностью). Назначить каждому ребру новый вес: $w(i,j) = \sum_{\forall m,n: Adj[i][m]=Adj[j][n]} [(k+1-m)(k+1-n)]$.

4. Для каждой вершины i вычислить общий вес ребер, инцидентных ей: $\mathrm{OBP}(i) = \sum_{j=0..|Adj[i]|-1} w(i, Adj[i][j])$.

5. Найти вершины-представители кластеров $\{i: \mathrm{OBP}(i) > \theta_1\}$.

6. Найти вершины, представляющие собой шум и выбросы $\{j: \mathrm{OBP}(j) < \theta_2\}$ и удалить их из графа.

7. Удалить все ребра с весом $< \theta_3$.

8. Удалить все вершины, которые не являются вершинами представителями, либо не соединены хотя бы с одной из них. Найти все связные компоненты графа. Они будут результирующими кластерами.

Время работы алгоритма (из-за шага 1) есть $O(|N|^2)$, где N – количество входных векторов, поскольку для многомерных данных нет общей методики быстрого определения k ближайших соседей. Для определенных предметных областей возможны методы оптимизации [48].

Центроид кластера – вектор, полученный усреднением всех векторов, вошедших в кластер. Пусть C_i – множество векторов кластера i, тогда $c_{ij} = \frac{\sum_{y_i \in C_i} y_{ij}}{|C_i|}$, где c_{ij}, y_{ij} – j-й элемент вектора центроида c_i и вектора y_i кластера C_i соответственно.

Используя в качестве климатической переменной давление на уровне моря, удалось получить временной ряд, коррелирующий с индексом Южного колебания в размере 0.78 [46, стр. 7] путем вычитания из центроида кластера 15 центроид кластера 20 (см. рис. 8). Аналогично разница центроидов кластеров 13 и 25 коррелирует в размере 0.81 с индексом Североатлантического колебания (North Atlantic Oscillation, NAO) – одного из важнейших климатических феноменов, оказывающее ключевое влияние на климат Европы и Северной Америки [43, стр. 536–539]. Индекс Североатлантического колебания определяется как нормализированная разница давлений на уровне моря между Ponta Delgada (столица Азорских островов) и Stykkisholmur (Исландия).

Существуют также индексы, основанные на температуре поверхности океана. Например, NINO1+2 вычисляется на основе температуре региона в районе 80° З.Д. – 90° З.Д. и 5° Ю.Ш. – 5°С.Ш, а NINO3.4 в 120° З.Д. – 170° З.Д. и 5° Ю.Ш. – 5°С.Ш [45, стр. 3].

Используя вместо давления температуру поверхности океана, было получено 107 кластеров (см. рис. 9).

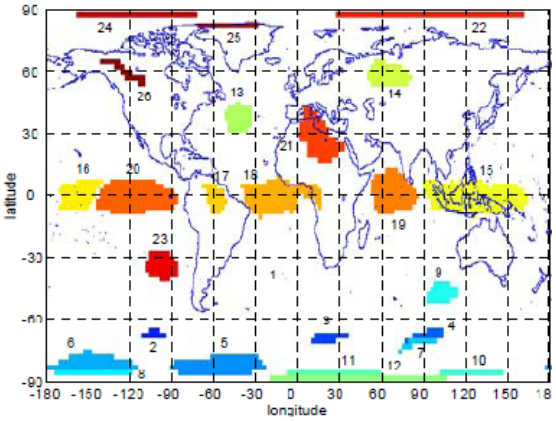

Рисунок 8. – Кластеры давления на уровне моря за 1982–1993 гг., полученные алгоритмом SNN [46, стр. 4, рис. 5]

Чтобы оценить влияние центроида кластера температуры на сушу, для него вычисляется сумма абсолютных значений его корреляций со всеми временными рядами ячеек суши.

С помощью описанного подхода, удалось получить кластеры (группа G0), центроиды которых которые коррелируют с индексами NINO1+2, NINO3, NINO3.4 и NINO4 в размере > 0.9 [45, стр. 7].

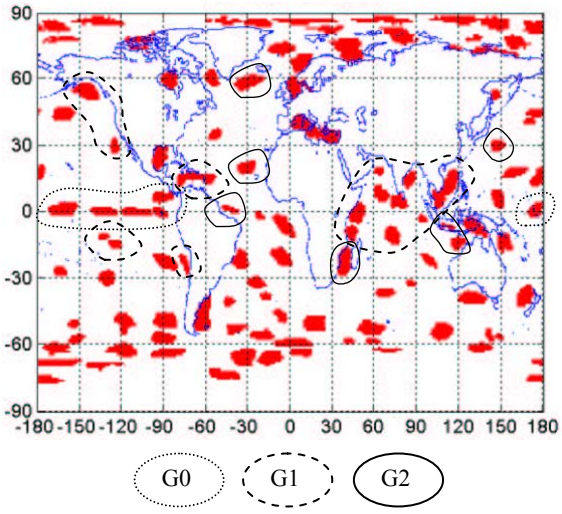

Рисунок 9. – Кастеры температуры поверхности океана за 1958–1998 гг., полученные алгоритмом SNN [45, стр. 6, рис. 7]. Разделены на 3 группы: G0, G1 и G3. Число кластеров–107

Также были получены альтернативные версии существующих индексов – центроиды кластеров группы G1, которые вероятно, относятся к тем же климатическим феноменам. Их центроиды более коррелируют с некоторыми областями суши, чем известные индексы.

Наконец, получены кластеры группы G3, которые слабо коррелируют с известными индексами, но имеют более высокую корреляцию с некоторыми областями суши, чем существующие индексы, а значит, могут представлять потенциально новые климатические феномены.

К ограничениям описанного подхода можно отнести кластеризацию только по одной переменной и то, что в основном коррелируют только экстремальные явления.

Например, корреляция центроидов кластеров 20 и 15, в момент проявления Эль-Ниньо высока и гораздо ниже в обычное время. При этом некоторые менее явные либо более кратковременные телеконнекции могут быть не обнаружены данным подходом.

Одним из решений может быть сравнение только определённых частей временных рядов. Также могут оказаться более адекватными подходы, основанные на событиях, например, ассоциативный анализ (описан в следующем разделе).

Еще одна особенность телеконнекций – мобильные кластеры, которые также затрудняют обнаружение климатических феноменов [51, стр. 3]. Например, NAO возникает через нерегулярные интервалы времени и точное расположение этого феномена изменяется от месяца к месяцу. Одним из решений по [51] может быть разработка алгоритмов поиска кластеров, которые изменяются во времени.

Ассоциативный анализ

Традиционно ассоциативные правила определяются на примере потребительской корзины (market basket) – набора товаров и услуг, приобретаемых на рынке, следующим образом [47, стр. 329–330; 52, стр. 258; 53, стр. 139; 54, стр. 207].

Транзакцией называется набор товаров, приобретенных покупателем за визит. Пусть Υ – множество всех транзакций, $count(Q) = |\{T: T \in \Upsilon, Q \subseteq T\}|$ – количество транзакций, в которых содержится набор товаров Q.

Ассоциативное правило – отношение вида $A \to B$, где A и B – наборы товаров, $A \neq \emptyset$, $B \neq \emptyset$, $A \cap B = \emptyset$. Ассоциативное правило характеризуется *поддержкой* (*support*), $sup(A \to B) = count(A \cup B)/|\Upsilon|$ и *достоверностью* (*confidence*), $conf(A \to B) = count(A \cup B)/count(A)$. Для некоторых порогов $minsup$ и $minconf$ правило $A \to B$ называют *часто встречающимся* (*frequent*), если $sup(A \to B) \geq minsup$, а если вдобавок $conf(A \to B) \geq minconf$ то и *сильным* (*strong*).

Например, если $A = \{\text{Масло}, \text{Молоко}\}$, $B = \{\text{Хлеб}\}$, $sup(A \to B) = 0.45$, $conf(A \to B) = 0.75$, то это можно интерпретировать так: «в 45% случаев покупатель приобретает масло, молоко и хлеб, причем, если покупатель приобретает масло и молоко, то в 75% случаев он также приобретает хлеб».

Ассоциативное правило вида $A \to B$ не выражает причинную связь, для утверждения существования которой необходимо знание причинно-следственных отношений. Напротив, оно свидетельствует о частом появлении элементов множества $A \cup B$ вместе.

В работе [55] впервые был выполнен ассоциативный анализ климатических данных. Рассматривались временные ряды температуры, количества осадков, солнечной радиации, а также NDVI, NPP, PET и FPAR (Fractional Intercepted Photosynthetic Active Radiation), измеренные в ячейках широтно-долготной сетки планеты.

Индекс FPAR вычисляется на основе NDVI. Аномально высокие значения FPAR означают, что растительность проявляет больше светособирающей фотосинтетической активности, чем обычно.

Наряду с изучением обычного хода природных процессов, климатологи заинтересованы в изучении явлений, которые являются отклонением от нормы. Поэтому из временных рядов выделяются аномальные события, определяемые как превышение либо падение значения климатической переменной выше либо ниже $\mu \pm 2\sigma$, где μ – математическое ожидание, а σ – среднеквадратическое отклонение. Будем обозначать эти события префиксами «Выс-» и «Низ-» соответственно.

Нас интересуют ассоциативные правила вида «для рассматриваемой территории характерны высокие значения PET наряду с низкими значениями Температуры, которые сопровождаются высокими значениями Солнечной радиации в 99.4% случаев».

При поиске ассоциативных правил, совокупность аномальных событий для региона представляется в виде транзакций с потребительскими корзинами. Главное преимущество такого подхода в том, что можно использовать существующие алгоритмы.

Приведенное выше правило можно записать в виде {Выс-PET, Низ-Температура} → {Выс-Солнечная рад}, $conf = 99.4\%$.

Пусть транзакцией будет набор аномальных событий $z_t = \{E_1, E_2, \dots, E_N\}$, которые имели место в момент времени t на интересующей нас территории (считается, что все эти события произошли одновременно). Тогда база данных транзакций будет состоять из всех z_t.

Алгоритм $Apriori$ использует простое свойство меры поддержки: если множество Q является часто встречающимся, то любое его подмножество $Q' \subseteq Q$ является также часто встречающимся. Действительно, Q' содержится, по крайней мере, во всех множествах, в которых содержится и Q, значит $count(Q') \geq count(Q)$.

Ниже представлено переработанное описание алгоритма на основе [47, стр. 332–349; 56, стр. 1–12]. Обозначим через $Q[i]$ элемент множества Q, который находится на -й позиции в лексикографически упорядоченной последовательности элементов Q.

1 $k = 1$
2 $F_1 = \{f: f \in \Upsilon, |f| = 1, count(f)/|\Upsilon| \geq minsup\}$
3 повторять
4 $C_{k+1} = \{X \cup Y:\ X, Y \in F_k,\ X[i] = Y[i],\ i = 1, \dots, k-1,\ X[k] \neq Y[k]\}$
5 для $\forall c \in C_{k+1}$
6 $K_c = \{s:\ s \subset c, |s| = k\}$
7 для $\forall s \in K_c$
8 если $s \notin F_k$
9 то удалить c из C_{k+1}
10 положить $count(c) = 0$ для $\forall c \in C_{k+1}$
11 для $\forall T \in \Upsilon$
12 $C_T = \{c: c \in C_{k+1}, c \subseteq T\}$
13 увеличить $count(c)$ на 1 для $\forall c \in C_T$
14 $F_{k+1} = \{c: c \in C_{k+1},\ count(c)/|\Upsilon| \geq minsup\}$
15 $k = k + 1$
16 до тех пор, пока $F_k \neq \emptyset$
17 Результат $= \cup F_k$

Рисунок 10. – Алгоритм $Apriori$

Вначале (строка 2) генерируются часто встречающиеся синглетоны. Каждая итерация алгоритма (строки 3–16) генерирует множество F_k – часто встречающиеся множества мощностью k (ЧВМ$_k$).

Алгоритм расширяет найденные ЧВМ$_k$ таким образом, чтобы не упустить ни одного ЧВМ$_{k+1}$ и при этом сгенерировать как можно меньше $k+1$ множеств, которые заведомо не являются часто встречающимися.

Для этого в строке 4 генерируются кандидаты в ЧВМ$_{k+1}$ следующим образом. Элементы множеств из ЧВМ$_k$ упорядочиваются в лексикографическом порядке и отыскиваются все пары ЧВМ$_k$, которые отличаются одним последним элементом. Кандидатами в ЧВМ$_{k+1}$ являются объединения найденных пар множеств.

Можно было бы расширять каждое ЧВМ$_k$ каждым ЧВМ$_1$, получая в итоге $|F_k| \times |F_1|$ кандидатов. Если множество $\{A, B, C\} \in$ ЧВМ$_3$, то обязательно также существуют $\{A, B\}, \{A, C\} \in$ ЧВМ$_2$. Поэтому нет необходимости расширять $\{A, B\} \in F_2$ множеством $\{C\} \in F_1$ либо объединять $\{A, B\}, \{B, C\} \in$ ЧВМ$_2$.

Описанный подход в разы уменьшает количество кандидатов $|C_{k+1}|$ в ЧВМ$_{k+1}$, которое существенно влияет на время выполнения остальной части алгоритма.

Можно отсечь некоторых кандидатов в ЧВМ$_{k+1}$ без обращения к базе данных транзакций. Для $Q \in$ ЧВМ$_{k+1}$ обязательно, чтобы все подмножества Q мощностью k (генерируются в строках 5–6), были ЧВМ$_k$ (проверяется в строках 7–9). Заметим, что $|K_c| = k + 1$, считая те два множества, объединением которых был получен кандидат.

Если все подмножества мощностью k кандидата Q в ЧВМ$_{k+1}$ являются ЧВМ$_k$, то это не гарантирует $Q \in$ ЧВМ$_{k+1}$. Для проверки оставшихся кандидатов, выполняется сканирование всей базы данных транзакций (строки 11–13) и подсчет количества транзакций, в которых содержатся кандидаты.

В строке 14 отбираются только те кандидаты, встречаемость которых удовлетворяет пороговому значению поддержки.

Алгоритм завершается, если не было найдено ни одного ЧВМ$_{k+1}$ (строка 16), а значит ЧВМ$_{k+i} = \emptyset$ для $\forall i > k + 1$ (гарантируется монотонностью меры поддержки). Результат работы алгоритма – все часто встречаются вместе наборы товаров.

Простым перебором всех товаров $X \in Q$ легко получить из набора $Q \in$ ЧВМ$_{|Q|}$ все правила вида $Q - X \rightarrow X$ при $|X| = 1$, проверяя при этом условие $minconf \leq conf(Q - X \rightarrow X) = count(Q)/count(Q - X)$. Случай $|X| > 1$ описан в [47, 54].

В результате были получены следующие ассоциативные правила (показаны первые 4 с наиболее высокой достоверностью) [55, стр. 9].

1. {Выс-PET, Выс-Осадки, Выс-FPAR, Выс-Температура}→{Выс-NPP}, c□nf = 100%.
2. {Выс-PET, Низ-Температура} → {Выс-Солнечная рад}, conf = 99.4%.
3. {Выс-PET, Выс-Осадки, Выс-FPAR} → {Выс-NPP}, conf = 98.6%.
4. {Низ-NPP, Низ-PET, Выс-Температура} → {Низ-Солнечная рад}, conf = 98.0%.

Используя вместо меры поддержки корреляцию *corr*(*A*,*B*) = *P*(*A*∪*B*) – *P*(*A*)*P*(*B*) / *sqrt*(*P*(*A*)(1 – *P*(*A*))P(*B*)(1 – *P*(*B*))), где *P*(*Q*) = *count*(*Q*)/|Y|, удалось получить такие правила:

1. {Низ-FPAR} → {Низ-NPP}, corr = 0.4327.
2. {Выс-FPAR} → {Выс-NPP}, corr = 0.4013 (см. рис. 11).
3. {Низ-Solar} → {Низ-PET}, corr = 0.2752.
4. {Низ-PET, Низ-FPAR} → {Низ-NPP}, corr = 0.1966.

Интересна закономерность, когда аномально высокие значения FPAR ведут к выше

обычным значениям NPP (см. рис. 11). Большинство регионов, которые проявляют такую закономерность, в основном соответствуют полузасушливым ежегодным пастбищам – типу растительности, которое способно к более быстрому извлечению пользы от периодически сильных осадков, чем леса.

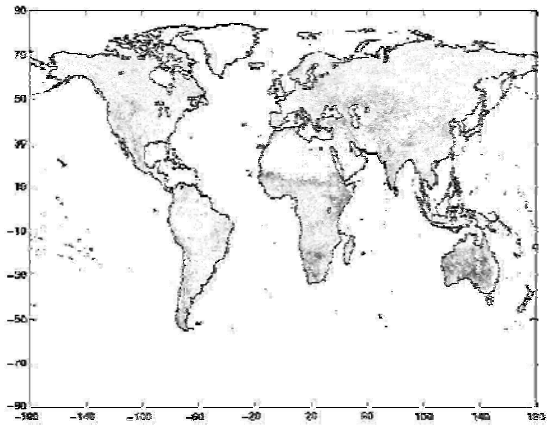

Рисунок 11. – Регионы, для которых {Выс-FPAR} → {Выс-NPP} [55, стр. 10, рис. 13]

Заключительные замечания

Когда исследователь в области компьютерных наук проводит эксперименты с алгоритмом, то он может изменить его входные параметры и повторить эксперимент еще раз. В климатологии такой характер познания невозможен. На сегодняшний день считается, что оптимальным способом познания окружающей среды является сбор максимального количества сведений о ней в единицу времени.

Метеорологические данные поступают со стационарных наземных метеорологических и аэрологических станций, островных станций в океане, погодных судов, судов добровольного наблюдения, различного вида подводных и поверхностных буев.

Наибольшее количество информации поступает со спутников, на борту которых установлены приборы дистанционного измерения температуры, влажности, рельефа, цвета и других параметров, позволяя узнавать о Земле из космоса больше, чем находясь на ней.

Например, спутник Национального управления США по исследованию океанов и атмосферы (NOAA) оснащён усовершенствованным радиометром с очень высоким разрешением (AVHRR) – 1.1 × 1.1 км. Прибор способен оценивать температуру поверхности океана и принимать излучения в видимом красном и инфракрасном областях спектра, позволяя рассчитывать индексы растительности, например NDVI.

Индекс NDVI позволяет узнать о типе растительности, почвы, и влаги в ней, характере землепользования, объеме биомассы.

Прибор AVHRR 4-х функционирующих спутников NOAA, имеет 10-битное квантование по 5 спектральным каналам и составляет полную картину земли за сутки, собирая таким образом почти 1.5 Гб данных в день.

Спутник NASA EOS Landsat 7 генерирует 7 Гб/день, а спутник Terra 194 Гб/день. В открытом доступе находится около 284 Терабайт данных, собранных спутниками NASA EOS [57].

Широтно-долготная 17-ти уровневая решетка с разрешением 2.5°×2.5° со значениями метеовеличин в узлах занимает 194.99 Гб и составляет всего 6.6% от общего объема архива повторного анализа NCEP/NCAR версии 1.

Учитывая, что скорость чтения с диска составляет около 75 Мб/сек, потребуется более 40 минут, чтобы только загрузить эту решетку с диска в последовательном режиме в оперативную память.

При таких объемах данных вычисление тривиального арифметического выражения, например, среднего, уже является далеко не тривиальной задачей.

Телеконнекции представляют собой одновременное изменение климата в разных далеко расположенных друг от друга географических районах. От поведения Североатлантического колебания зависит климат Северной Америки и Европы, а от Эль-Ниньо жизни и имущество тысяч людей Южной Америки и Австралии.

Наиболее раннее упоминание об Эль-Ниньо приходится на конец 19 в., когда перуанские моряки окрестили этим именем появление у берега необычно теплой воды на Рождество. Сегодня учет этого феномена является ключевым при сезонном прогнозе климата.

Для описание телеконнекций климатологи разрабатывают климатические индексы. Например, индекс Южного колебания, который коррелирует с периодами и степенью Эль-Ниньо, доступен с 1876 г. и представляет собой разность атмосферного давления на уровне моря между Дарвином (Австралия) и Таити.

Несмотря на кажущуюся зрелость и полноту данных о телеконнекциях, до сих пор ведется поиск новых индексов и телеконнекций.

Дело в том, что до эры дистанционных спутниковых измерений, приходилось пользоваться данными стационарных островных метеостанций.

Официальной методики вычисления конкретных индексов, признаваемых большинством исследователей и организаций нет [34, стр. 50].

С доступностью глобальных данных об океане высокого разрешения и открытых архивов повторного анализа, начали появляться дискуссии об адекватности используемых индексов: скорее всего, существуют другие участки океана, на основе которых можно построить

лучшие индексы. Поэтому стали применятся дополнительные индексы, например, NINO1+2, NINO3.4, которые рассчитываются на основе температуры определенных участков океана, при этом индекс Южного колебания стал утрачивать свою популярность [34, стр. 50–51].

В связи с этим, степень корреляции новых индексов для Эль-Ниньо и других климатических феноменов с индексом Южного колебания не свидетельствует о степени адекватности новых индексов. Сам индекс Южного колебания служит лишь описанием феномена Эль-Ниньо, существуя сам по себе.

Сегодняшний объем данных не позволяет вручную найти все полезные закономерности. В связи с этим, был выполнен поиск телеконнекций с помощью кластеризации. Вся территория океана автоматически, но при этом осмысленно разделяется на области (кластеры) с относительно однородным климатическим поведением. Центроиды найденных кластеров представляют собой временные ряды, в среднем характеризующие поведение соответствующих им областей.

Одни, из полученных таким образом, центроиды представляют собой уже известные климатические индексы океана и служат подтверждением правильности предложенного подхода, другие являются альтернативой известным индексам, которые имеют лучшую прогнозную способность для некоторых регионов, а третьи потенциально представляют собой новые климатические феномены.

Спутниковые данные позволили больше узнать о характере растительного покрова и его реакции на изменчивость климата.

С помощью алгоритмов ассоциативного анализа были обнаружены взаимоотношения между аномальными значениями метеорологических величин и индексов растительности (превышение либо падение значения выше либо ниже заданного порога).

Ассоциативные правила представлялись в виде транзакций с потребительскими корзинами, что позволило использовать существующие алгоритмы поиска ассоциативных правил, например *Apriori*.

Полученные таким образом правила представили интерес для климатологов и требуют дальнейшего исследования.

Однако используемый подход приводит к плотным матрицам транзакций и, следовательно, занимает существенное время.

Также при использовании стандартных мер важности ассоциативных правил не удается получить все интересные и потенциально полезные закономерности. Например, при использовании корреляции вместо меры достоверности были получены другие важные закономерности.

Проанализированный подход может быть расширен поиском трендов в обнаруженных ассоциативных правилах.

Территория Украины также находится под влиянием телеконнекций. При этом особое воздействие на Украину оказывает Североатлантическое колебание [58, 59]. Проанализированные в этой статье методы кластеризации и поиска ассоциативных правил могут найти новые потенциально полезные взщаимосвязи между крупномасштабными процессами и региональным климатом Украины. Это позволит использовать их для своевременного прогноза паводок, заморозков, засух и других чрезвычайных климатических ситуаций на территории Украины.

В 2001 г. группа исследователей в области компьютерных наук университета Миннесоты (США) получили грант НАСА для проведения первого в своем роде исследования климата с помощью интеллектуального анализа данных. Их цель «помочь ученым в области наук о Земле в их усилиях лучше понять» климатическую систему планеты не просто не утратил актуальности за прошедшие десять лет, но и получил новое развитие благодаря совершенствованию способов мониторинга природы нашей планеты.

Выводы

Впервые в одной статье проанализированы данные и методы их интеллектуального анализа для исследования окружающей природной среды.

Наиболее познавательным источником об окружающей среде являются данные дистанционного зондирования Земли, которые позволяют узнать из космоса больше, чем находясь на самой планете.

Разработка систем поддержки принятия решений на основе ДЗЗ – наиболее активная область исследований климата.

В Украине развивается система Geo-UA. Целями системы будут оценки возникновения чрезвычайных ситуаций, анализ и прогноз окружающей среды, выработка рекомендаций относительно рационального природопользования и решение другие стратегически важных управленческих задач [60, слайд 28].

По перечисленным задачам существует довольно разреженное множество методов интеллектуального анализа данных ДЗЗ. Это обусловлено тем, что до недавнего времени их разрешение не позволяло детально проанализировать интересующие аналитиков объекты (например, отдельные сельскохозяйственные поля и городские водоемы).

Большие объемы данных (сотни терабайт), актуальность задач поддержки принятия управленческих решений и недостаток методов интеллектуально анализа данных дистанцион-

ного зондирования Земли определяют высокопроизводительные вычисления и методы обнаружения изменений (change detection) [61] наиболее актуальными для дальнейшего исследования.

Поскольку в контексте программы Geo-UA Украина в этом году (2010) выведет на орбиту собственный спутник дистанционного зондирования Земли Сич-2 [62] с высоким пространственным разрешением (7 метров), то Донецкий национальный технический университет способен сделать существенный вклад в развитие программы устойчивого развития страны благодаря новым высокопроизводительным ресурсам (кластер из 200 ядер по 2.8 ГГц).

Литература

1. Что мы понимаем под погодой? // El. resource. URL: http://meteoprog.ua/blogs/mavis/2008/09/05/what_is_wheather/ (10.10.2010).
2. Data Mining: Overview // El. resource. URL: http://ocw.mit.edu/NR/rdonlyres/Sloan-School-of-Management/15-062Data-MiningSpring2003/E738593F-0DF6-47B7-9171-A39839F160AE/0/Lecture1Slides.pdf\ (10.10.2010).
3. Data Mining: Definition from Answers.com // El. resource. URL: http://www.answers.com/topic/data-mining (10.10.2010).
4. Zhizhin M. et al., Parallel storage, mining and visualization of environmental data archives, ER Symposium, 2009.
5. Ganguly A. and Steinhaeuser K., Data Mining for Climate Change and Impacts, ICDM-SSTDM, 2008.
6. IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon S. et al. (eds.)]. Cambridge University Press, NY, USA, 996 pp.
7. Бєловодський В.М. и Соавторы. Звіт про науково-дослідницьку роботу М2-08 «Изучение возможности разработки модели краткосрочного прогноза погоды для Донецкого региона», Донецький національний технічний університет, кафедра комп'ютерних систем моніторингу, 2008.
8. Зверев А.С. Синоптическая метеорология, 2-е изд., перераб. и доп. Л.: Гидрометеоиздат, 1977 – 712 с., ил.
9. Гордин В.А. Из чего делают прогноз погоды // El. resource. URL: http://edu.mccme.ru/Project/OL/chelife1.htm (10.10.2010).
10. Хейс Дж. Всемирная служба погоды сегодня, Бюллетень ВМО, 57(1), стр. 8–16, 2008.
11. Моура А.Д. Вклад ВМО в ГЕОСС и ГЕОНетКаст, Бюллетень ВМО, 55(4), 2006.
12. Система мониторинга окружающей среды в Донецкой области - Метеопоказатели - Графики // El. resource. URL: http://www.omos.org.ua/graph_meteo.php (10.10.2010).
13. NCDC: Online Climate Data Directory // El. resource. URL: http://lwf.ncdc.noaa.gov/oa/climate/climatedata.html (10.10.2010).
14. Cofiño A.S. et al. Implementation of data mining techniques for meteorological applications, Realizing Teracomputing. W. Zwieflhofer and N. Kreitz, eds., World Scientific, pp. 215–240, 2005.
15. Geo-Data: The World Geographical Encyclopedia / John F. McCoy, project editor. -- 3rd ed., ISBN 0-7876-5581-3, 717 pp., 2003.
16. АРМ «Метеоролога». АО «Специальные системы связи» // El. resource. URL: http://www.ssc.com.ua/index.php?option=com_content&task=view&id=22&Itemid=15 (10.10.2010).
17. Создание национальной сети перадчи гидрометеорогических данных (АСПД). АО «Специальные системы связи» // El. resource. URL: http://www.ssc.com.ua/index.php?option=com_content&task=view&id=38&Itemid=46 (10.10.2010).
18. Программно аппаратный комплекс метеорологической телесвязи «Бриз» , АО «Специальные системы связи» // El. resource. URL: http://www.ssc.com.ua/index.php? option=com_content&task=view&id=17&Itemid=17 (10.10.2010).
19. Бюро Экономического Анализа :: Общее объявление о торгах // El. resource. URL: http://www.beafnd.org/ru/projects/project_rosgidromet/notice_haggle/NHMP_GPN_2006/ (10.10.2010).
20. ЛАНИТ модернизирует метеорологическую сеть Росгидромета // El. resource. URL: http://job.lanit.ru/article/225 (10.10.2010).
21. Бахурел П. и Соавторы, Усвоение данных об океане в прогностической системе Меркатор Океан, Бюллетень ВМО, стр. 144–151, 2007.
22. Калинин Н.А., Толмачева Н.И. Космические методы исследований в метеорологии, учебник для студентов высших учебных заведений, обучающихся по специальности «Метеорология», ГОУВПО «Пермский государственный университет», 2005.
23. GIS-Lab: NDVI – [теория] и практика // El. resource. URL: http://gis-lab.info/qa/ndvi.html (10.10.2010).
24. Measuring Vegetation (NDVI & EVI): Feature Articles // El. resource. URL: http://earthobservatory.nasa.gov/ Features/MeasuringVegetation/measuring_vegetation_2.php (10.10.2010).
25. Primary Production, Photosynthetically Active Radiation (PAR) and Light Use Efficiency (LUE) // El. resource. URL: http://www.ccpo.odu.edu/SEES/veget/class/Chap_4/4_6.htm (10.10.2010).
26. Биологическая продуктивность // El. resource. URL: http://forest.geoman.ru/forest/item/f00/s00/e0000219/index.shtml (10.10.2010).
27. Evapotranspiration // El. resource. URL: http://en.wikipedia.org/wiki/Evapotranspiration (10.10.2010).
28. Хабиб Ш. и Соавторы. Наблюдения за Землей из космоса на благо общества, Бюллетень ВМО, 57(1), стр. 22–28, 2008.
29. USDA Crop Explorer: Global Crop Condition and Commodity Production Analysis from the USDA/Production Estimates and Crop Assessment Division (PECAD) // El. resource. URL: http://gcmd.gsfc.nasa.gov/ Keyword-Search/Metadata.do?Portal=GCMD&KeywordPath=Parameters%7CCLIMATE+INDICATORS%7CRefine+By+Locations%7CCONTINENT%7CEUROPE%7CEASTERN+EUROPE%7CUKRAINE&OrigMetadataNode=GCMD&EntryId=USDA0557&MetadataView=Full& MetadataType=0&lbnode=mdlb1(10.10.2010).
30. Kalnay E. et al. The NCEP/NCAR 40-Year Reanalysis Project, Bull. Amer. Meteor. Soc., 77, 437–472, 1996.
31. CISL RDA: ds090.0 Home Page // El. resource. URL: http://dss.ucar.edu/datasets/ds090.0/ (10.10.2010).
32. Kalnay E. Atmospheric modeling, data assimilation and predictability, Cambridge University Press, ISBN-13 978-0-511-07784-5, 369 pp., 2003.
33. Kanamitsu M. et al. NCEP–DOE AMIP-II Reanalysis (R-2), Bull. Amer. Meteor. Soc., 1631–1643, 2002.
34. Huug van den Dool. Empirical Methods In Short-Term Climate Prediction, Oxford University Press, ISBN 0-19-920278-8, 215 pp., 2007.
35. O'Lenic E. A. et al. Developments in operational long-range climate prediction at CPC, Weather and Forecasting, 23, 496–515.
36. Погода, климат и воздух, которым мы дышим, Бюллетень ВМО, Том 58(1), Январь 2009.
37. Коррe С. et. al. Периоды сильной жары: угрозы и ответные меры, Серия «Здоровье и глобальное изменение окружающей среды», № 2, ВОЗ, 2005.

38. Нешиба С. Океанология. Современные представления о жидкой оболочке Земли: пер. с англ. – М.: Мир, 1991. – 414 с., ил.
39. van Oldenborgh G. J. et al. Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over that last 15 years?, 2005.
40. Conlan R. and Service R., El Nino and La Nina: Tracing the Dance of Ocean and Atmosphere // El. resource. URL: www.nationalacademies.org/opus/elnino (10.10.2010).
41. NASA World Wind Java Demo Applications and Applets // El. resource. URL: http://worldwind.arc.nasa.gov/ java/demos/ (10.10.2010).
42. Сэр Гильберт Т. Уолкер. Climate Variability and El Nino, 1924.
43. Oliver J. E. (editor). Encyclopedia of World Climatology, Springer, ISBN-10 1-4020-3266-8 (e-book), 874 pp., 2005.
44. Monthly Southern Oscillation Index // El. resource. URL: ftp://ftp.bom.gov.au/anon/home/ncc/www/sco/soi/soiplaintext.html (10.10.2010).
45. Steinbach M. et al. Discovery of Climate Indices using Clustering, KDD 2003, 2003.
46. Steinbach M. et al. Data Mining for the Discovery of Ocean Climate Indices, The Fifth Workshop on Scientific Data Mining (2nd SIAM International Conference on Data Mining), 2002.
47. Tan P. N., Steinbach M., Kumar V. Introduction to Data Mining, Addison-Wesley, ISBN 0-321-32136-7, 2005.
48. Ertoz L. et al. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data , SIAM International Conference on Data Mining (SDM '03), 2003.
49. Ertoz L. et al. A New Shared Nearest Neighbor Clustering Algorithm and its Applications, Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining, 2002.
50. Ertoz L. et al. Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach, Clustering and Information Retrieval, Kluwer Academic Publishers, 2003.
51. Steinbach M. et al. The Application of Clustering to Earth Science Data: Progress and Challenges, 2005.
52. Shekhar S., ChawlaS. Spatial Databases: A Tour, Prentice Hall, ISBN 013-017480-7, 2003.
53. Барсегян А.А., Куприянов М.С. Степаненко В.В., Холод И.И., Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.: ил.
54. Agrawal R., Imielinsky T., Swami A. Mining Association Rules Between Sets of Items in Large Databases, In Proc. ACM SIGMOD Intl. Conf. Management of Data, 207–216, Washington, DC, 1993.
55. Tan P. et al. Finding Spatio-Temporal Patterns in Earth Science Data, KDD Workshop on Temporal Data Mining, 2001.
56. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules, IBM Almaden Research Center, 1994.
57. EOSDIS - Wikipedia, the free encyclopedia // El. resource. URL: http://en.wikipedia.org/wiki/EOSDIS (10.10.2010).
58. Семёнова И.Г. Циркуляционные условия атмосферы в периоды катастрофических летних паводков на Дунае, Вестник Одесского государственного экологического университета, вып. 6, стр. 103–109, 2008.
59. Ефимов В.А., Ивус Г.П. Хаджи-Страти Е.Д., Кумулятивные процессы и атмосферная телеконнекция на территории Украины, Научные работы УкрНГГМИ, вып. 256, стр. 155–165, 2007.
60. Федоров О.П. Исследование Земли из космоса: опыт Украины и ближайшие планы, 2007 // El. resource. URL: http://d33.infospace.ru/d33_conf/2007_pdf/plenar/fedorov.pdf (10.10.2010).
61. Boriah S., Kumar V., Steinbach M., Potter C., Klooster S. Land Cover Change Detection: A Case Study, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.
62. Подробности.UA Украинский спутник "Сич 2" выведут на орбиту в октябре // El. resource. URL: http://podrobnosti.ua/technologies/2010/04/13/678730.html (10.10.2010).