Data Min Knowl Disc (2012) 24:311–324
DOI 10.1007/s10618-011-0239-5

EDITORIAL

Introduction to data mining for sustainability

Katharina Morik · Kanishka Bhaduri · Hillol Kargupta

Received: 2 September 2011 / Accepted: 12 September 2011 / Published online: 12 October 2011

1 The need for sustainability

According to Brundtland Commission of the United Nations, *sustainability* can be defined as "capacity to endure the needs of today's population without jeopardizing the ability of the future generations to meet their own needs". Sustainability implies resource consumption with little internal or external adverse impact. A system or a process is sustainable if its input and output have little adverse impact on its environment. A system that is not sustainable often leads to the failure (sometimes catastrophic) of the system itself or other systems in its environment. For example, a banking system is sustainable when its financial transactions lead to little or no adverse effect on itself or other economic entities it interacts with. A domestic article such as a light bulb is sustainable when its production and operation are reliable and result in little adverse impact on its environment. For humans, sustainability has long term effects on three important sectors: environmental, economic, and social. Earth system science pleads for global sustainability. The goal is to deliver the knowledge which allows to reduce

K. Morik (✉)
TU Dortmund University, Computer Science Faculty, Artificial Intelligence Group, Dortmund, Germany
e-mail: katharina.morik@tu-dortmund.de

K. Bhaduri
Mission Critical Technologies Inc., NASA Ames Research Center, Intelligent Data Understanding, Moffett Field, CA, USA
e-mail: Kanishka.Bhaduri-1@nasa.gov

H. Kargupta
Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, MD, USA
e-mail: hillol@cs.umbc.edu

global environmental risks. In this special issue, data mining techniques are presented that explore and analyze environmental spatio-temporal data or help to design and operate better sustainable systems.

The introduction is structured as follows. First we describe the particular challenges and tasks of data mining for sustainability (Sect. 2). An overview of data mining in the earth sciences indicates work towards the identified tasks (Sect. 2.1). Data mining for the design and control of sustainable systems is illustrated by transportation issues and energy-awareness (Sect. 2.2).

As a service section, we include an overview of organizations that provide data mining with data repositories (Sect. 3). This might inspire others to apply their algorithms to the wealth of environmental data.

We conclude with an overview of this issue's articles (Sect. 4).

2 Data mining tasks and problems

The earth sciences acquire and manage a large variety of data, based on satellites, aircrafts, ships, buoys, autonomous underwater vehicles. Diverse sensors deliver the data, ranging from accelerometer, seismic sensors, radiometers, measures of light level, temperature, humidity, soil moisture to cameras, video cameras. Sensor networks report to a central base station and may share their data with other nodes which allows to adapt their measuring behavior (for an overview of such projects cf. Hart and Martinez (2006)). Engineering approaches measure technical processes at various, distributed locations in order to minimize the resource consumption. In addition to the large amount of data and their distribution, the measure process needs to be understood, managed and controlled. The data collections of environmental and engineering approaches to sustainability are challenging data mining in various ways. The *challenges* are in particular the following:

Scalability Most of the data sets are tremendously large. Since scalability has been a key challenge of data mining from its very beginning, most of the algorithms are evaluated w.r.t. scalability. The listed repositories in Sect. 3 offer real-world data for testing the scalability of algorithms.

Integration Several data sets measure the same event. Mapping of the measurements that are distributed over several sources to events is a difficult task, since the sources may use different time, space, bandwidth granularities.

Distributed data mining Even if the observations of different data sets can easily be mapped, pushing them into one central repository is prohibited by their large size and communication cost. Hence, distributed data mining is on demand.

Real-time prediction Timely warning is necessary, for instance, in the context of disaster management and risk assessment. The online application of an offline learned model may be sufficient, but often a streaming data or online algorithm is used.

Spatio-temporal data The possibly high-dimensional data sets are organized into spatial and temporal neighborhoods. The relation between these two orderings must be taken into account by the mining algorithms. Spatio-temporal data mining chal-

lenges are described in Ganguly and Steinhaeuser (2009) together with a case study on heat wave data.
Understandability Many tools have been developed that display the data in different views and allow for interactive analysis (cf. the above listed sites of organizations in earth science). Mapping found patterns on Google maps, for instance, helps to understand the data. Since the complexity of several data generating principles often exceeds human capabilities of understanding, decomposing the data or factorization helps to single out aspects. Understandability is stressed by Schwabacher et al. (2007).

The goal of earth sciences is the understanding of global and local processes in geosciences, climate sciences, and biosciences. A better understanding is supposed to be a prerequisite for better sustainability. Several institutes work on simulations that model the earth system. The integration of more specific models for weather, land, oceans, and atmosphere is supposed to enhance predictions (Ferraro et al. 2003). Models can also be built using statistical and data mining methods. In a study on machine learning for sustainability, the term "ecosystem informatics" has been used for stressing computational aspects (Dietterich et al. 2009). Here, we roughly distinguish the following areas of an *environmental approach* to sustainability:

Disaster management Risk analysis concerning Tsunami, earth quake, volcanic eruption, landslides, fire outbreak is necessary in order to take appropriate measures on time.
Climate Weather forecast has a long tradition and is constantly put forward regarding the interaction of atmosphere, ocean, and land. Weather forecast is crucial for agricultural management and other economic enterprises. Nowadays, the debate on climate change asks if the weather processes are stationary, or not.
Natural resources Water, agricultural, and forestry management aim at keeping the biological systems diverse and productive. Biodiversity has become a hot topic at an international level.

Another approach to sustainability is to enhance the management of human consumption of resources. As an *engineering approach*, it aims at controlling processes such that natural resources are conserved. Data are acquired about economic processes, traffic and transport processes, social behavior, urban planning, and individual lifestyles. Also computation and its energy consumption is under investigation. A view of data mining towards distributed sensor measurements also for sustainability can be found in the book (May and Saitta 2010).

Data analysis contributes to both, the environmental and the engineering approach. Given the storage of measured or simulated data, the main *tasks of data mining for sustainability* are:

Data exploration and visualization Several tools are available from the earth science web sites, e.g., from NCAR, there is the Man computer Interactive Data Access System (McIDAS) and the package GEMPAK for meteorological data. Such tools ease the data access. Principal component analysis and singular value decomposition and simple statistics are often used in data inspection and as a preprocessing step for further analysis. Visual analytics is a key to a first analysis (Andrienko et al. 2010).

Interactive exploration of spatio-temporal data (Andrienko et al. 2009) contribute to the first step of data analysis, as well.

Pattern mining Unsupervised techniques for detecting frequent patterns or high-density regions, both, in time and space, deliver findings that are of interest in their own right or form the basis for further analysis.

Anomaly detection Since the interesting patterns are rare, the mining task can be considered outlier detection. Without the negative connotation of outliers, we may also be named local patterns as opposed to global patterns (Morik et al. 2005; Furnkranz and Knobbe 2010). An overview of methods for anomaly detection is Chandola et al. (2009).

Prediction and forecast Prediction and forecast in the earth sciences is often based on simulations. Data mining tasks are classification and regression. Sometimes, these unsupervised methods are a prerequisite for anomaly detection.

We may think of data mining for sustainability in terms of a matrix with the topics of the environmental and engineering approaches as columns and the tasks of data mining as rows. All entries in this matrix face at least some of the challenges. In the following, we illustrate the field of data mining for sustainability along these lines.

2.1 Environmental approaches

Data exploration and visualization In the area of biodiversity and biological resource management, the standard is to use simple statistical techniques. The impact of soil parameters for plant productivity or succession has been investigated. For instance, 147 plots were chosen in nitrogen limited soil grassland and the number of plant species in each plot was controlled. Then, scores for the productivity of plants were correlated and, hence, shown that more diverse plant communities are more productive (Tilman et al. 1996). Another study inspects soil parameters using statistics and simulation with respect to plant succession, recolonization mechanisms, functional group interactions, mycorrhizal diversity and function, and food web analysis (Blanke et al. 2007). Such statistical methods are only possible for very small data sets, both in the number of examples and features. Currently, earth sciences are busy with sensing, storing, and managing very large data volumes. If these data volumes are to be used, scalable methods from data mining are required.

Analysis is made in a close human computer interaction, where the computer merges measurements from diverse sources and displays them and the user interprets the images. For instance, at the Center for Satellite-based Crisis Information (DLR, Germany), the risk assessment of regions regarding Tsunami is developed (see Fig. 1). However, this close interaction with human experts makes current data volumes a problem. Hence, data mining is in demand. An approach to exploit data mining methods for better human–computer collaboration in data analysis can be found in this issue (Thurau et al. 2011).

Pattern mining One of the first data mining contributions to climate investigations is Kumar et al. (2001). A sequence of earth shots is preprocessed using Fourier transform and Singular Value Decomposition. A monthly Z-score is extracted as a useful feature. Moving average over the time series removes seasonal variations. Frequent

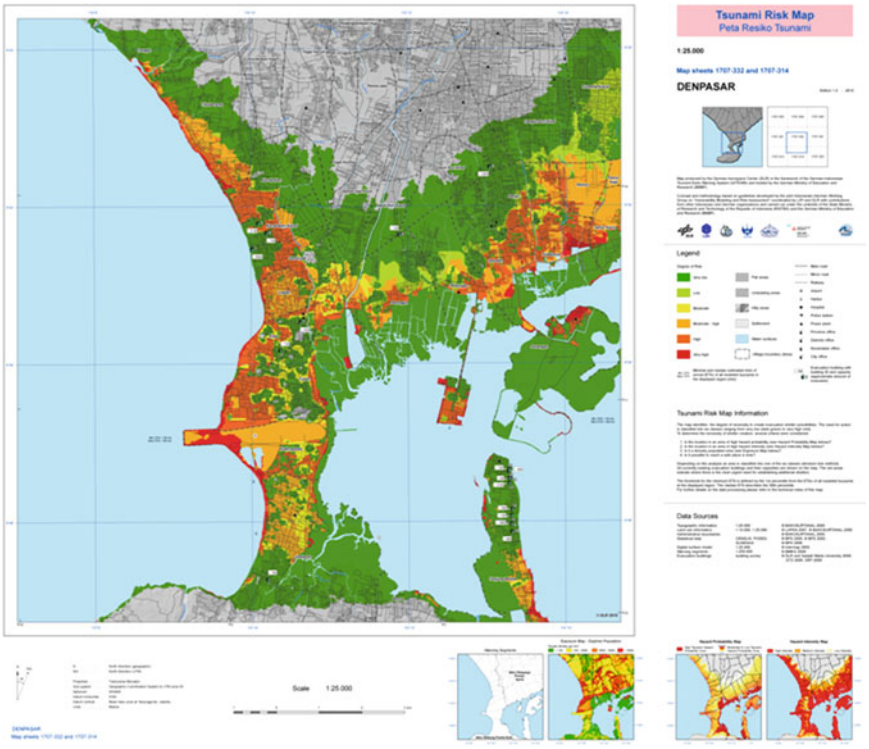

Fig. 1 The Tsunami risk assessment mapped on the map of Denpasar (Bali) (http://www.zki.dlr.de/system/files/media/filefield/image/risk_map.jpg), a result of the German-Indonesian Tsunami Early Warning System project

set mining and K-Means Clustering is applied so that separate regions of land and ocean are constructed that show seasonal patterns and associations between land and ocean. The seasonal patterns are contrasted by anomaly regions.

Steinbach et al. (2003) frame the problem of determining summaries of time series. Grouping together regions of similar behavior has been performed carefully by geo science, developing indices such as NIN 1+2, which expresses the average sea surface temperature anomaly off the coast of Peru. Such hand-crafted indices are correlated with climate anomalies, e.g., heavy rainfall or drought. Using Principal Component Analysis or Singular Value Decomposition builds indices more easily. Based on Sea Surface Temperature data or Sea Level Pressure, several indices have been formed. However, these methods find orthogonal and strong signals, hence, not all relevant indices are detected. In contrast, clustering is capable of detecting patterns that fall into the same subspace or overlapping subspaces. The Shared Nearest Neighbor (SNN) algorithm redefines similarity between points in terms of how many nearest neighbors the two points share. This results in core points around which clusters build. The advantage is that the number of clusters is automatically determined (the number of core points). Clusters of low density in uniform regions can be found. Based on SST data, an ocean index was developed by SNN. These clusters can be interpreted and

correlate with known indices. It was shown that SNN clusters outperformed SVD results. A follow-up paper investigates the correlations of ocean indices and land temperature (Boriah et al. 2004).

Anomaly detection Using time series techniques for change detection regarding seasonal effects (Boriah et al. 2008), validation of fire events using the California fire database. The variable vegetation index (EVI) indicating "greenness" of vegetation is based on the Moderate Resolution Imaging Spectroradiometer (MODIS). It is given for the time span February 2000 until January 2006. The task is to detect level changes. Recursive Merging Algorithm on time series of 380,400 locations:

1. compare $year_i$ and $year_{i+1}$ and store the distance—this handles the seasons;
2. recursively merge the most similar years until only one annual cycle is left.
3. Output the change score for each location: ration of maximal and minimal distance over the years.
4. Manually inspecting the few events with the highest score.

Scalability of outlier detection has been investigated in data mining (cf. e.g., Bay and Schwabacher (2003), Ramaswamy et al. (2000)). A new method builds on distributed data mining in order to achieve scalability together with handling the distributed storage of data. Detecting anomalies in satellite data from the MODerate-resolution Spectroradiometer (MODIS) imaging is investigated concerning the impact of parallel processing (Bhaduri et al. 2010). The approach scales well, because the 1class-SVM (Schoelkopf et al. 2001) is trained in parallel on each of the features. An outlier with respect to one feature is considered an outlier also in the high-dimensional model. Then, the global model is adjusted using the found local outliers. The procedure is very accurate (98.33–99.79%) compared to the non parallel version and is communication efficient. Anomalies due to diverse causes could successfully be detected in the satellite images of California.

Analyzing the daily global air temperature from heterogeneous sensors from 1950 to 1999 is to point at spatial, temporal, and spatio-temporal outliers (Das and Parthasarathy 2009). K-Means and Principal Component Analysis are used to find regions which are similar w.r.t. climate trends. This unsupervised learning is a prerequisite for the following prediction task. Support Vector Machines and Neural Networks are used for predicting climate behavior.

Prediction and forecast The prediction of which crop farmers will decide to grow on their land, allows to better manage water resources. Given socio-economic data from a farmers' survey, including the farmers' assessment of the soil and use of irrigation, a decision tree is learned which predicts the farmers's choice (Ekasingh et al. 2005). Accordingly, the irrigation water of the district can be planned. Water management is also concerned in a study of stormwater tanks where the fill level is predicted using genetic programming (Flasch et al. 2010).

The effect of emissions from fire on air quality and climate is investigated in Mazzoni et al. (2006). Given measurements of the Terra Multi-angle Imaging SpectroRadiometer (MISR),[1] smoke plumes are to be identified and distinguished from clouds or other types of aerosols. The data are enriched by the MODerate-resolution

[1] Data come from NASA Data Active Archive Center DAAC.

Spectroradiometer (MODIS) that is located on the same Terra spacecraft as is MISR but delivers data about fire.[2] In these data together with meteorological data plumes are to be discovered and their direction and injection height to be retrieved. Within the large amount of data there are only few smoke plumes. As a rough first pruning, a standard feature of MODIS classifies each pixel as fire, cloud, water, or land. All non-fires are no longer considered, rejecting about 82% of the data. Now, the corresponding pixels of fire regions in MISR images are more carefully inspected. A few scenes are manually labeled as cloud, smoke, dust, land, water, and ice or snow. From these few examples, the support vector machine trains a classifier for the six classes. Since recall is far more important than precision here, the classifier is biased in favor of smoke. The shape of a plume is determined by an additional brightness threshold. For a detected plume its height is calculated, enhanced by a wind correction. The distance from the plume source is related to the plume's height.

Forecasting storm, particularly mesocyclones, has been investigated using neural networks, clustering, and image processing based on weather data and simulations (Li et al. 2008). The main achievement is the connection between data acquisition and data mining, in that data can trigger the data analysis. Another contribution to storm prediction using machine learning methods can be found in this issue (Stojanova et al. 2011).

2.2 Engineering approaches

A more engineering type approach considers the control of processes such that resources are conserved. The engineering approach can be applied to scientific as well as commercial data. Reducing the footprint of our activities and saving energy has become a hot topic. An important area is transportation and traffic. We illustrate this application by one representative, the system MineFleet (Kargupta et al. 2010). MineFleet is a commercially available data mining system for commercial fleets. MineFleet analyzes high throughput data streams onboard the vehicle, generates the analytics, sends those to the remote server over the wide-area wireless networks and offers them to the fleet managers using stand-alone and web-based user-interface. It has several capabilities which aim at better fleet management and reduce green house emissions. The data collected from each vehicle is analyzed locally and then significant components of the data are transmitted to the Minefleet server. Minefleet consists of the following key modules:

- *Predictive health monitoring* This method processes the diagnostic data coming from the vehicle data bus and correlates with the maintenance data to build a predictive model. However, since a car may operate in many regimes, instead of building one model, a separate one is built for each regime. A number of health tests are executed against this current model and an alarm is raised whenever the current data does not fit the model.
- *Fuel consumption analysis* Minefleet system has the capability of computing the fuel economy of a vehicle/fleet, performing trend analysis of various kinds, and

[2] Data come from Earth Resources Observation and Science Data Center.

correlating them with various vehicle and driver performance parameters. Since fuel combustion is one of the main sources of green house emissions, any small reduction in fuel savings helps in reducing global warming.
– *Driver behavior monitoring* This module of Minefleet identifies the speeding, braking, idling characteristics of the driver and use that for driver retraining policy execution. It also assigns performance measures to the drivers based on various characteristics and identify outlier drivers. Finally it can identify unusual maintenance operations caused by suboptimal driver performance.
– *Emission monitoring* Greenhouse gas (GHG) emissions that contribute to climate change are a global problem. Although future concentrations, damages and costs are unknown, it is widely recognized that major emissions reduction efforts are needed. Of the four primary GHG under scrutiny, carbon dioxide (CO_2), and the need to lower carbon emissions in general, is of paramount concern. It is estimated that transportation activities are responsible for approximately 25% to 30% of total U.S. GHG emissions, with the on-highway commercial truck market accounting for over 45% of transportation GHG. However, the transportation sector emissions remain almost entirely unaddressed with respect to GHG and CO_2 reduction. Minefleet can measure the emissions data in real-time, correlate that with the vehicle performance and traffic data using advanced statistical and machine learning-based techniques such as clustering, predictive modeling, correlation analysis and eigen analysis. These analytics can be used to offer a new generation of decision support tools to develop fleet and greenhouse gas emissions management policies.

Computing itself consumes energy and, hence, becomes subject to acquiring data of resource consumption. Green computing analyzes the data of computing centers (cf., e.g., Marwah et al. 2009). At a finer granularity, program codes and operating systems are inspected. Energy-aware codes (e.g., Lorenz et al. 2001) and memory organizations (e.g., Steinke et al. 2002) are investigated. Measurements of different architectures and platforms are compared regarding their energy consumption. First studies apply data analysis methods. For instance, detecting usage patterns in system call data allows to adapt the operating system of ubiquitous systems (including mobile phones) such that energy is saved (Fricke et al. 2010).

3 Organizations, projects, data repositories, and tools

In this section we present a brief listing of some projects and groups who are performing active research pertaining to the field of sustainable systems (environmental, economic, and social). We also present several datasets related to these projects and encourage users to apply advanced knowledge discovery algorithms to the data.

1. *Institute name* University of Minnesota, USA
   url (http://climatechange.cs.umn.edu/index.php,http://gopher.cs.umn.edu/)
   Description Climate change is one of the most critical challenges facing the human society today. However, there is considerable uncertainty in understanding these changes due to the limited capabilities of the existing low fidelity physics-based models of the Earth system. On the contrary, huge volumes of observational data

is available from satellite and ground-based sensors recording the atmospheric, oceanic, and terrestrial processes, and physics-based climate model simulations. The goal of this project is to apply advanced data mining techniques in order to discover the not yet known climate phenomenon from such massive volumes of data.

Another project focuses on global warming. This project is looking at the following 4 sub areas: (1) Data Mining for the Discovery of Ocean Climate Indices, (2) Discovery of Patterns in the Global Climate System, (3) Finding Spatio-Temporal Patterns in Earth Science Data, and (4) Clustering Earth Science Data: Goals, Issues and Results.

Data links Investigating climate change uses data from a variety of sources. The National Oceanic and Atmospheric Administration (NOAA) (http://www.noaa.gov/) publishes many climate datasets. One of the most popular among them is the NCEP/NCAR Reanalysis 1 dataset which is available at http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html. Also this project uses the MODIS (or Moderate Resolution Imaging Spectroradiometer) dataset which is a key instrument aboard the Terra and Aqua satellites. Terra's orbit around the Earth is timed so that it passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. Terra MODIS and Aqua MODIS are viewing the entire Earth's surface every 1–2 days, acquiring data in 36 spectral bands and in resolutions of 250 m, 500 m and 1 km. There are many data products derived from MODIS observations which describe features of the land, oceans and the atmosphere that can be used for studies of processes and trends on local to global scales. The link to data download can be found at: http://modis.gsfc.nasa.gov/data/.

Global warming is investigated using various products of the MODIS dataset http://modis.gsfc.nasa.gov/data/. In particular it uses the Enhanced Vegetation Index (EVI) distributed through the Land Processes Distributed Active Archive Center (LP DAAC https://lpdaac.usgs.gov/lpdaac/products/modis_products_table).

2. *Project name* Terrestrial Observation and Prediction System (TOPS): Developing ecological nowcasts and forecasts by integrating surface, satellite and climate data with simulation models at Ecological Forecasting Lab at NASA Ames Research Center, USA

   url http://ecocast.arc.nasa.gov/

   Description Ecological forecasting is very challenging primarily due to the availability of several sources of heterogeneous information which need to be combined in order to predict correctly. Lack of infrastructure for integrating a variety of modeling tools, information technologies, and ground and satellite data sets that could serve the diverse needs of eco-hydrological community has been one key issue is such cohesive prediction systems. This project aims at developing an integrated system called the Terrestrial Observation and Prediction System (TOPS) which is a data and modeling software system designed to seamlessly integrate data from satellite, aircraft, and ground sensors with weather/climate and application models. The goal is to provide nowcasts and forecasts of ecological conditions and events.

   Data links This project generates and uses a number of datasets. All the data products are available at the NASA Earth Exchange (NEX) website at: https://c3.nasa.

gov/nex/. This is a great resource for data mining practitioners who are interested in working in the area of climate/earth sciences.

3. *Institute name* Sustainability Research Institute, School of Earth and Environment, University of Leeds, Leeds, UK
   url http://www.see.leeds.ac.uk/research/sri/
   Description The Sustainability Research Institute explores a wide range of issues including climate change, energy, transport, water, resource use, land use, conservation, cities and communities, business and lifestyles. The research within the SRI is based largely on the environmental social sciences and draws upon aspects of geography, sociology, politics, planning, economics, management, development studies and science and technology studies. Some of the active areas of research are: (1) sustainable development and environmental change, (2) environmental policy, planning and governance, (3) ecological and environmental economics, (4) business, environment and corporate responsibility, and (4) sustainable production and consumption.
4. *Project name* Transportation sustainability research center, UC Berkeley, CA USA
   url http://www.tsrc.berkeley.edu/About/index.html
   Description The TSRC mission is to conduct research, educate, and engage in outreach to improve understanding of the economic, environmental, and social aspects of transportation systems. Reducing vehicle emissions by initiatives related to sustainable transportation is a major goal of TSRC. They have several active projects: (1) Advanced Vehicles and Fuels, (2) Energy and Infrastructure, (3) Goods Movement, (4) Integrated Land Use, (5) Transit and Travel Connections, and (5) Mobility for Special Populations.
   Data links Some of the datasets for the projects are available at http://www.tsrc.berkeley.edu/Resources/data.html.
5. *Institute name* The Center for BioEnergy Sustainability (CBES) at Oak Ridge National Laboratory Oak Ridge, Tennessee, USA
   url http://www.ornl.gov/sci/ees/cbes/
   Description The Center for BioEnergy Sustainability (CBES) does active research on sustainability of biomass production for conversion to biofuels and bio-based products. Its goal is to use data and analysis to understand the sustainability of current and potential future bioenergy production and distribution and to identify approaches to enhance bioenergy sustainability. This in turn would help decision makers in taking informed decision on bioenergy consumption and production.
   Data links The link http://www.ornl.gov/sci/ees/cbes/publications.shtml contains several publications by the group which contains information about some of the biomass data sources.
6. *Institute name* Centre for Integrated Sustainability Analysis at The University of Sydney, Australia
   url http://www.isa.org.usyd.edu.au/index.html
   Description The centre for Integrated Sustainability Analysis develops leading-edge research and applications for environmental and broader sustainability issues, bringing together expertise in environmental science, economics, technology, and social science.

7. *Institute name* The International Research Institute for Climate and Society, New York, USA
   url http://portal.iri.columbia.edu/portal/server.pt
   Description The IRI uses a data driven/science-based approach to enhance society's capability to understand, anticipate and manage the impacts of climate in order to improve human welfare and the environment, especially in developing countries.
   Data links The IRI/LDEO Climate Data Library contains over a large number of datasets from a variety of earth science disciplines and climate-related topics. The data can be downloaded from: http://iridl.ldeo.columbia.edu/index.html.

There are several other datasets available to data mining researchers which we enumerate below:

1. The Earth Simulator Center at the Japan Agency for Marine-Earth Science and Technology publishes publishes several datasets from their simulation studies which are available at http://www.jamstec.go.jp/esc/download/index.en.html.
2. Several datasets are available at the Max Planck Institute for Meteorology (Germany). These datasets comprise both observed and model generated data available at: http://www.mpimet.mpg.de/en/wissenschaft/datensaetze.html. The link http://www.mpimet.mpg.de/en/links.html contains information for several institutes who publish datasets on climate, observed as well as simulated.
3. The German Research Centre for Geosciences (GFZ) offers tools and several data collections: GEOFON Global Seismic Monitor (seismology); Geomagnetic Observatory Niemegk (Earth magnetic field); ICDP Scientific Drilling Database (results from scientific drilling); ICGEM—International Centre for Global Earth Models (geodesy and gravimetry); ISDC—Information System and Data Center (geodetic satellite missions); WSM—World Stress Map (tectonic stress vectors), available at: http://www.gfz-potsdam.de/portal/gfz/Services/Forschungsdaten. The Centre participates in the Earth Observation System (EOS), which focuses on ice and ocean, processes of the land surface, and disaster management. EOS offers data as well as tools and papers at: http://helmholtz-eos.dlr.de/start_en.htm.
4. The NASA Goddard Institute for Space Studies (USA) combines analysis of comprehensive global datasets with global models of atmospheric, land surface, and oceanic processes. It publishes a wealth of data products for many different projects. All the data is available for download at: http://data.giss.nasa.gov/.
5. The UCAR COSMIC Data Analysis and Archive Center (CDAAC) provides data for climate, space weather and geodetic research. The data is collected from six identical micro satellites, each carrying measurement instruments. Data access is free, but users must submit a data agreement. The data is available at: http://cosmic-io.cosmic.ucar.edu/cdaac/.
6. The VAMOS Ocean-Cloud-Atmosphere-Land Study (VOCALS) is an international program the major goal of which is to develop and promote scientific activities leading to improved understanding of the coupled ocean-atmosphere-land system on diurnal to inter-annual timescales. They publish several datasets all of which can be downloaded from http://data.eol.ucar.edu/master_list/?project=VOCALS.

7. The Climatic Research Unit at School of Environmental Sciences, University of East Anglia, UK develops methods to improve scientific understanding in both past, present and future climate history and its impact on humanity. The center distributes several datasets which can be downloaded from: http://www.cru.uea.ac.uk/cru/data/.
8. The Intergovernmental Panel on Climate Change (IPCC) publishes several climate datasets which can be downloaded at: http://www.ipcc-data.org/obs/index.html.
9. The Royal Meteorological Society publishes temperature and precipitation data of central England and other parts of UK, which can be downloaded from http://www.rmets.org/weather/climate/datasets.php.
10. The Joint Institute for the Study of the Atmosphere and Ocean (JISAO) at University of Washington, USA hosts several datasets and scripts to read these datasets. The data is available from http://jisao.washington.edu/data/.

4 Overview of this issue's contributions

The articles of this issue discuss several data mining tasks, namely exploration, mining of periodic patterns, tracing patterns over time, and prediction. Accordingly, a variety of methods is proposed covering matrix factorization, clustering, classification, and regression. The methods face spatio-temporal data and their models are to be understood by domain experts or just citizens. Possibly due to the public nature of many environmental data most of the papers contribute to natural resources and climate issues. The engineering approach is frequently investigated within companies and often constitutes a competitive advantage so that data and methods are not easily published. However, one of the articles in this issue presents a method of energy saving in air traffic (Srivastava 2011).

1. The paper "Descriptive Matrix Factorization for Sustainability" proposes a new method for *data exploration* based on matrix factorization that enhances the *understandability*. Understanding data is considered a process of detecting influential components, here seen as basis vectors. Constraints guarantee that the basis vectors reside on actual data points and usually at the very extreme. These properties make the results better understandable than those of other decomposition methods. The challenge of *scalability* is answered by a linear-time algorithm. For illustration, *climate* data of the USA and world-wide *electricity consumption* are explored.
2. The paper "Mining Periodic Behaviors of Object Movements for Animal and Biological Sustainability Studies" introduces *pattern mining* of time periods that faces the complexity of *spatio-temporal data*. Raw trajectories are not well suited for the analysis. Places, which are visited several times by several animals are selected as reference spots. For each place, a sequence of movements is represented. Periods within these sequences are then determined using Fast Fourier Transform and autocorrelation. The application is in the area of *natural resources*, namely the management of animal population—a topic found important in ecology (Cross et al. 2005).

3. Another aspect of *spatio-temporal patterns* is handled by the paper "Tracing Evolving Subspace Clusters in Temporal Climate Data". As opposed to tracing animals in the previous paper, where the same stork is observed at different places and the behavior of this animal is to be described, this paper investigates behaviors that are given by similar attribute values. Climate situation are not objects in the usual sense but are constituted by certain sets of attribute values. Tracing such *climate situations*, here oceanographic data, is conducted using a *subspace clustering* method.
4. The paper "Estimating the risk of fire outbreaks in the natural environment" uses *spatio-temporal data* sampled from areas in Slovenia and enriched by a simulation of weather forecasts. The goal of fire prevention is an instance of saving *natural resources*. The data mining task is *prediction*. Several classification algorithms are evaluated on the environmental data.
5. The paper "Greener aviation with virtual sensors: A case study" represents the engineering approach, here the area of energy saving in *traffic*. Data from 84 aircrafts are gathered from more than 1900 flights. The task of data analysis is *anomaly detection*. An ensemble of regressions allows to determine such aircrafts that consume more energy than expected. The method moves beyond threshold methods for the comparison of actual and average consumption. The analysis aims at determining points where the aircraft construction should be enhanced.

References

Andrienko GL, Andrienko NV (2009) Interactive cluster analysis of diverse types of spatiotemporal data. SIGKDD Explor 11(2):19–28

Andrienko GL, Andrienko NV, Demsar U, Dransch D, Dykes J, Fabrikant SI, Jern M, Kraak M-J, Schumann H, Tominski C (2010) Space, time and visual analytics. Int J Geogr Inf Sci 24(10):1577–1600

Bay SD, Schwabacher M (2003) Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proceedings of the international conference on knowledge discovery and data mining KDD. pp 29–38

Bhaduri K, Das K, Votava P (2010) Distributed anomaly detection using satellite data from multiple modalities. In NASA conference on intelligent data understanding (CIDU'10). pp 109–123

Blanke V, Schulze B, Gerighausen U, Küster S, Rothe R, Schulze H, Sineriz M (2007) The power of regeneration: lessons from a degraded grassland. Restor Ecol 15(2):307–311

Boriah S, Simon G, Naorem M, Kumar V, Steinbach M, Potter C, Klooster S (2004) Predicting land temperature using ocean data. In: Proceedings of the knowledge discovery in databases KDD

Boriah S, Kumar V, Steinbach M, Potter C, Klooster S (2008) Land cover change detection: a case study. In: Proceedings of the knowledge discovery in databases KDD

Chandola V, Berjee A, Kumar V (2009) Anomaly detection—a survey. ACM Comput Surv 41(3)

Cross PC, Lloyd-Smith JO, Johnson PLF, Getz WM (2005) Dueling timescales of host movement and disease recovery determine invasion of disease in structured populations. Ecol Lett

Das M, Parthasarathy S (2009) Anomaly detection and spatio-temporal analysis of global climate systems. In: SensorKDD

Dietterich TG (2009) Machine learning in ecosystem informatics and sustainabiility. In: Proceedings of 21st IJCAI. AAAI

Ekasingh B, Ngamsomsuke K, Letcher RA, Spate J (2005) A data mining approach to simulating farmers' crop choices for integrated water resources management. J Environ Manag 77:315–325

Ferraro R, Sato T, Brasseur G, DeLuca C, Guilyardi E (2003) Modeling the earth system. In: Proceedings of the international geoscience and remote sensing symposium

Flasch O, Bartz-Beielstein T, Davtyan A, Koch P, Konen W, Oyetoyan TD, Tamutan M (2010) Comparing spo-tuned gp and narx prediction models for stormwater tank fill level prediction. In: Proceedings of the IEEE congress on evolutionary computation, Barcelona. pp 1–8

Fricke P, Jungermann F, Morik K, Piatkowski N, Spinczyk O, Stolpe M (2010) Towards adjusting mobile devices to user's behaviour. In: Workshop mining ubiquitous and social environments (MUSE) at ECML PKDD

Fürnkranz J, Knobbe AJ (2010) Guest editorial: global modeling using local patterns. Data Min Knowl Discov 21(1):1–8

Ganguly AR, Steinhaeuser K (2009) Data mining for climate change and impacts. In Proceedings of IEEE international conference on data mining (ICDM)workshops. pp 385–394

Hart JK, Martiez K (2006) Environmental sensor networks: a revolution in the earth system sciene? Earth Sci Rev 78:177–191

Kargupta H, Sarkar K, Gilligan M (2010) Minefleet: an overview of a widely adopted distributed vehicle performance data mining system. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp 37–46

Kumar V, Steinbach M, Tan PN, Klooster S, Potter C, Torregrossa A (2001) Mining scientific data: discovery of patterns in the global climate system. In Proceedings of the joint statistical meetings of the American Statistical Association

Li X, Plale B, Vijayakumar N, Ramachandran R, Graves S, Conover H (2008) Real-time storm detection and weather forecast activation through data mining and events processing. Earth Sci Inform. pp 49–57

Lorenz L, Leupers R, Marwedel P, Dräger T, Fettweis GP (2001) Low-energy dsp code generation using a genetic algorithm. In: ICCD '01

Marwah M, Sharma R, Shih R, Patel C, Bhatia V, Mekanapurath M, Velumni R, Velayudhan S (2009) Data analysis, visualization and knowledge discovery in sustainable data centers. In: Compute

May M, Saitta L (eds) (2010) Ubiquitous knowledge discovery volume 6202 of Lecture Notes in Artificial Intelligence. Springer

Mazzoni D, Logan JA, Diner D, Kahn R, Tong L, Li Q (2006) A data-mining approach to associating MISR smoke plume heights with MODIS fire measurements. Remote Sens Environ 107:138–148

Morik K, Boulicaut J-F, Siebes A (eds) (2005) Local pattern detection, international seminar, Dagstuhl Castle, Germany, April 12–16, 2004, Revised Selected Papers volume 3539 of Lecture Notes in Computer Science. Springer

Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Rec 29(2):427–438

Schölkopf B, Platt JC, Shawe-Taylor JC, Smola AJ (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13(7)

Schwabacher M, Langley P, Potter C, Klooster SA, Torregrosa A (2007) Discovering communicable models from earth science data. In: Dzeroski S, Todorovski L (eds) Computational discovery of scientific knowledge. Springer, Berlin pp 138–157

Srivastava AN (2011) Greener aviation with virtual sensors: a case study. Data Min Knowl Discov

Steinbach M, Tan PN, Kumar V, Klooster S, Potter C (2003) Discovery of climate indices using clustering. In Proceedings of KDD

Steinke S, Wehmeyer L, Lee B, Marwedel P (2002) Assigning program and data objects to scratchpad for energy reduction. In: DATE '02: Proceedings of the conference on Design, automation and test in Europe page 409, Washington, DC, USA, IEEE Computer Society

Stojanova D, Kobler A, Dzersoki S (2011) Estimating the risk of fire outbreaks in natural environment. Data Min Knowl Discov

Thurau C, Kersting K, Wahabzada M, Bauckhage C (2011) Descriptive matrix factorization for sustainability—adopting the principle of opposites. Data Min Knowl Discov

Tilman D, Wedin D, Knops J (1996) Productivity and sustainability influenced by biodiversity in grassland ecosystems. Nature 379:718–720