Data Mining and Scientific Data

Authors: Raymond, Ben, Watts, David J., Burton, Harry, and Bonnice, Jeremy

Source: Arctic, Antarctic, and Alpine Research, 37(3) : 348-357

Published By: Institute of Arctic and Alpine Research (INSTAAR), University of Colorado

URL: https://doi.org/10.1657/1523-0430(2005)037[0348:DMASD]2.0.CO;2

Arctic, Antarctic, and Alpine Research, Vol. 37, No. 3, 2005, pp. 348–357

Data Mining and Scientific Data

Ben Raymond†*
*David J. Watts**
*Harry Burton** *and*
*Jeremy Bonnice**

*Australian Government, Department of the Environment and Heritage, Australian Antarctic Division, Channel Highway, Kingston 7050, Australia
†ben.raymond@aad.gov.au

Abstract

Data mining—the discovery of previously unknown information from a large collection of individual data sources—is becoming increasingly popular for scientific data archives. We describe an approach to data mining that uses spatial, temporal, and type constraints to obtain a broad list of data that are potentially related to a data set of interest. Tree- and spline-based multivariate regression and classification techniques are then used to identify functional relationships between the data. Expert knowledge is used to constrain and guide the model building and evaluation process.

We demonstrate the approach by identifying relationships between indicators in a state of the Antarctic environment reporting database. Analyses of the fuel usage of electrical generators and boilers at Australia's Davis station yielded fuel usage dependencies on air temperature and wind speed that were in good accordance with known physical processes. The phenomenon of periodic haul-outs of large numbers of leopard seals on Macquarie Island was related to anomalies in regional sea ice cover and sea surface temperature.

Introduction

The Australian Antarctic Data Centre (AADC) was established in 1995 in order to provide a coordinated facility for managing the scientific data collected by the Australian Antarctic scientific program. Creating an aggregated collection of data such as this makes data maintenance more efficient, provides users with simplified access to data in a consistent format, and can assist in saving data that might otherwise be lost over time in individual scientists' notebooks and computer files. A data center can also provide a critical mass of information from which previously unknown patterns and relationships can emerge. The active pursuit of this process—data mining—is widespread in the corporate sector, finding application in fields such as insurance risk analysis (Apte et al., 1999), analysis of customer transaction databases (Agrawal et al., 1993), and of event sequences in telecommunications fault databases (Mannila et al., 1997). In recent years, data mining techniques have become increasingly applied to scientific data. Examples include searching for specific patterns in very large collections, such as astronomical data (Ng et al., 1998; Rocke and Dai, 2003) and global satellite observations (Potter et al., 2003), and using regression rules to relate spatial and temporal dependences between climatic and vegetation data (Schwabacher and Langley, 2001).

Data mining has been given varying definitions in the literature (Friedman, 1997; Ramakrishnan and Grama, 2001). Here we use the term to describe the entire process of discovering knowledge in databases, including such aspects as database design and the validation of discovered knowledge. Data mining is complementary to traditional statistical analyses of scientific data. While the formulation of scientific hypotheses has conventionally followed the observation of physical phenomena, the observation of numerical properties of previously collected data can also provide this stimulus (Crawford and Crawford, 1996). Hypotheses formed in this manner can be tested using existing data, or where this is inadequate, by further physical experiment or observation. Data mining therefore has a role in the strategic planning of scientific research.

The boundary between data mining and conventional statistical methods is not well defined (Glymour et al., 1996; Friedman, 1997; Hand, 1999). The distinction is generally drawn on the grounds of complexity: data mining typically operates on very large data sets with many variables, to which classical statistical methods often do not scale well. The processes of data mining and data maintenance are also tightly linked, so that database design and interaction must be carefully considered (Chaudhuri, 1998; Mannila, 2000). Data mining is also commonly applied to data originally collected for some other purpose and can be considered to be a secondary analysis (Mannila, 2000).

Antarctic science is in many ways a prime candidate for data mining. While remote sensing can be used to obtain measurements of some environmental variables, the direct acquisition of data from the Antarctic is very expensive and logistically difficult. Data mining offers a means of extracting maximum scientific value from expensive data. The diversity of Antarctic terrestrial ecosystems is relatively low (Bergstrom and Chown, 1999) and so, despite the wide range of types of data collected, many of these data might be interrelated. Further, sampling locations tend to be concentrated in those areas that are relatively easy to access (generally, near to Antarctic stations). Disparate projects often collect data from the same location, giving a variety of data at relatively few sampling sites. These characteristics are broadly true of data collected from any harsh, remote environment. Issues related to Antarctic data mining are therefore likely to be applicable to data from other cold regions.

We present two examples of data mining analyses of a small subset of the AADC's holdings. The aim of the investigation was to establish a methodology for identifying functional relationships among a variety of data sources within the AADC.

Methods

The principal database used in this investigation was the System for Indicator Management and Reporting (SIMR, accessible online at http://www.aad.gov.au/soe). The SIMR was designed specifically to facilitate Australian Antarctic State of the Environment reporting (Belbin et al., 2003). At the heart of the SIMR is a set of environmental indicators (a partial list appears in Table 1). Each indicator is a variable that measures an aspect of environmental conditions, a pressure applied to the environment by human activities, or a response that has been initiated to minimize an environmental pressure. These indicators

1523-0430/05 $7.00

TABLE 1

A partial list of State of the Environment indicators, sorted by theme. C, P, and R denote that the indicator reflects a condition, pressure, or response (some indicators may reflect more than one of these). See http://www.aad.gov.au/soe for more information.

Indicator	Type
Theme: Atmosphere	
Daily broadband ultraviolet radiation observations using biologically effective UVR detectors	C
Highest monthly air temperatures at Australian Antarctic stations	C
Lowest monthly air temperatures at Australian Antarctic stations	C
Monthly mean air temperatures at Australian Antarctic stations	C
Monthly mean atmospheric pressure at Australian Antarctic stations	C
Monthly mean lower stratospheric temperatures above Australian Antarctic stations	C
Atmospheric concentrations of greenhouse gas species	CP
Daily records of total column ozone at Macquarie Island	CP
Theme: Biodiversity	
Fecundity and pup growth in fur seal colonies on Macquarie Island	C
Regional populations of Adelie penguins in the vicinity of Casey, Davis and Mawson	C
The presence or absence of vascular plant species in two defined areas of Heard Island	C
Annual catch in tonnes of marine species harvested in Australian Antarctic and sub-Antarctic waters	CP
Species and number of species killed, taken or interfered with or disturbed in the Antarctic and the sub-Antarctic for the purpose of scientific research	P
Theme: Coasts and Oceans	
Fast ice thickness at Davis and Mawson	C
Mean sea level for the Antarctic region	C
Theme: Human Settlements	
Medical consultations per 1000 person years	C
Quality of potable water at Australian Antarctic and subantarctic stations	C
Annual tourist ship visits and tourist numbers	P
Biological Oxygen Demand (BOD) of wastewater discharged from Australian Antarctic stations	P
Monthly electricity usage at Australian Antarctic stations	P
Monthly fuel usage of the generator sets and boilers	P
Monthly incinerator fuel usage of Australian Antarctic stations	P
Monthly total of fuel used by vehicles at Australian Antarctic stations	P
Station and ship person days	P
Suspended solids content of wastewater discharged from Australian Antarctic stations	P
Total potable water consumption at Australian Antarctic stations	P
Volume of wastewater discharged from Australian Antarctic stations	P
Amount of waste incinerated at Australian Antarctic stations	PR
Waste returned to Australia	PR
Number of expeditioners undergoing environmental education	R
Resources committed to environmental issues	R
Theme: Land	
Water levels of Deep Lake, Vestfold Hills	C
Station footprint for Australian Antarctic stations	P
The number of permits issued for entry into Antarctic Specially Protected Areas (ASPAs) in the Australian Antarctic Territory	P
Resources committed to heritage expertise	R
The number and area of protected areas in the Australian Antarctic and sub-Antarctic jurisdiction	R

provide an ongoing, objective record of aspects of the state of the Antarctic environment. The SIMR indicator set encompasses a wide diversity of information, including measurements of atmospheric and marine conditions, biodiversity, and direct human impacts on the environment (see Table 1). As of October 2004, there were 43 publicly accessible indicators in the SIMR, although the indicator set is dynamic and continuing to evolve. This database was chosen for investigation because the broad diversity of information contained within it is representative of that of Antarctic data in general. Some of the indicators in the database are widely studied and well understood processes, and some are not. The better studied indicators provide a testing ground on which to prove data mining techniques, before tackling the more difficult data held by the AADC.

The model fitting and selection aspects of the data mining process can be described in terms of regression and classification. The goal is to model the dependence of a response variable y on a set of explanatory variables $x_1 \ldots x_n$ (often called independent variables). The response y may be categorical, in which case the problem becomes one of classification. Not all of the explanatory variables will be relevant. The objective is therefore to identify which of the explanatory variables are relevant to the response variable and to suggest the forms of the relationships between them. The naive approach of trying each explanatory variable in turn rapidly becomes infeasible when higher-order terms are included to allow for nonlinear behavior. There are a large number of established methods for multivariate regression and variable selection—see, for example, Miller (2002) and Hastie et al. (2001). Here, we use two techniques that simultaneously select relevant variables and construct the model: classification and regression trees (Breiman et al., 1984) and multivariate adaptive regression splines (Friedman, 1991). Both techniques are nonlinear and use relatively efficient search strategies to locate relevant explanatory variables from those available. Regression trees use a recursive set of if . . . then rules to construct an approximation to the response variable. Trees are fast to build, and can provide an intuitive illustration of the relationships among the variables provided that the tree does not become too large. Trees have a considerable advantage over many multivariate techniques in that there is no need to choose a specific measure of association. This can be particularly appealing when working with categorical or ordinal data. Missing data can be handled by finding surrogate predictors, capitalizing on shared information among explanatory variables. Trees can also be used for classification purposes. However, trees have several drawbacks. Regression trees can produce models that are discontinuous, which can make model interpretation difficult. Trees also do not work well with small data sets (Hastie et al., 2001) and can often perform poorly when modeling additive interactions between explanatory variables. An example of a regression tree is shown in Figure 1.

The technique of multivariate adaptive regression splines (MARS; Friedman, 1991) offers some advantages when compared to the tree methodology. MARS first constructs a set of building blocks: basis functions that are linear over part of the input range of an explanatory variable, and zero over the remainder of its range. These functions are iteratively added to the model to construct a progressively more

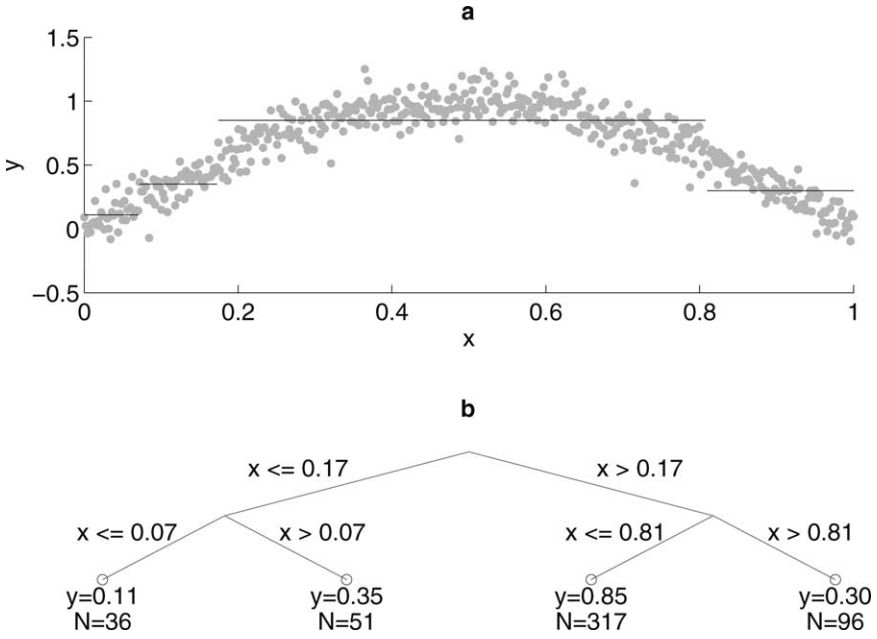

FIGURE 1. An example of a regression tree. The target data were drawn from a noisy half-sinusoid and are shown as the gray points in (a). The regression tree approximation is the solid line, and the tree itself is shown in (b).

detailed approximation to the response variable. An example is shown in Figure 2. In its simplest form, MARS produces a model that is piecewise linear with respect to the explanatory variables. Allowing products of two or more basis functions to be added to the model introduces higher-order terms and interactions among variables. The method is more computationally demanding than trees, but is better able to handle additive structures (Friedman, 1991) and is likely to offer better performance with small data sets. MARS can also be used for classification (e.g., Hastie et al., 1994).

Data mining requires a high degree of interaction with the users (Crawford and Crawford, 1996; Hand, 1999; Mannila, 2000). Here, the data miner (a scientist with a background in statistical methods) was the principal driver of the process. Other scientists with expertise in relevant disciplines were consulted and participated throughout the process. The role of the discipline experts was to guide the variable selection and model building process, as described below, and to provide critical evaluation of the discovered models. The data mining process is outlined in broad terms by the flowchart in Figure 3. The first step of the process was to ensure that the problem was in a suitable form for regression analysis. In many cases this required a transformation of the response variable. Other preprocessing included the removal of outliers and seasonal variations. Environmental and other data often display a strong seasonal component that may not be scientifically important; it is the other variations (e.g., long-term trends) in the data that are often of interest. Once a suitable form of the response variable had been established, a semi-automated search (as described below) for potentially relevant explanatory variables was carried out. The search included other indicators in the SIMR database, as well as other data held by the AADC, such as automatic weather station data from each of the four Australian stations, and remotely

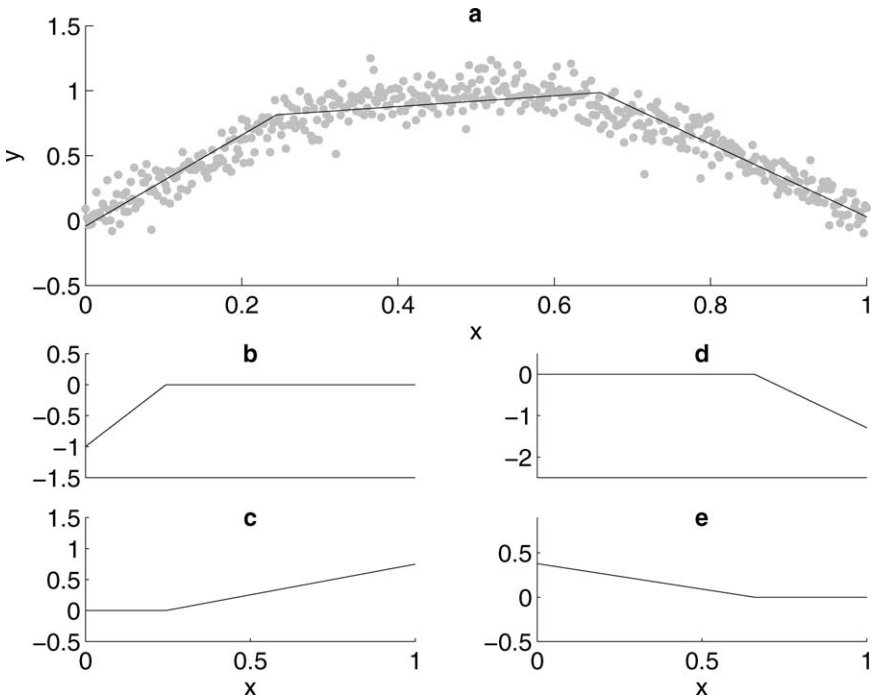

FIGURE 2. An example of a multivariate adaptive regression spline (MARS) model. The target data were drawn from a noisy half-sinusoid and are shown as the gray points in (a). The MARS approximation is the solid line. The MARS model is made up of a sum of piecewise linear basis functions (b–e). Compare the accuracy of this model to that obtained using a regression tree (Fig. 1).

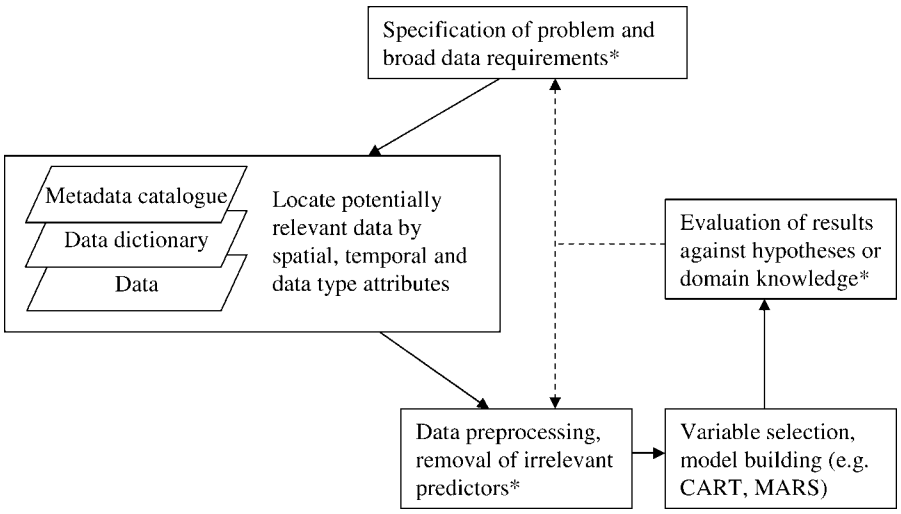

FIGURE 3. A flowchart of the data mining process. Asterisks denote steps that involve interaction with a domain expert in order to constrain and guide the process. CART = classification and regression trees; MARS = multivariate adaptive regression splines.

sensed environmental data (Reynolds et al., 2001; Cavalieri et al., 1999; Smith and Reynolds, 2003). Each data source held by the AADC has an associated metadata record: a complete description of the data, including sampling methods, spatial and temporal coverage, and the details of the personnel responsible for maintaining the data. The data type is also documented, and linked to a data dictionary that defines the measurement units and acceptable ranges of values for that data type. This metadata was used to automatically construct a shortlist of potentially relevant explanatory data, using spatio-temporal coverage and data type as initial search criteria. Subsets of data with broad spatio-temporal coverage (such as global sea surface temperature data sets) were used where appropriate. The spatial extents of these subsets were chosen by the expert. This initial search was quite general and often yielded explanatory variables with no conceivable connection to the response variable. Irrelevant explanatory variables were removed from the shortlist at the discretion of the relevant expert. Overly aggressive removal of apparently irrelevant explanatory variables was discouraged because it could remove variables with a previously unknown connection to the response variable.

In many cases the temporal granularity of the explanatory and response variables differed (e.g., a yearly response variable but monthly explanatory data). Shorter-term data were transformed to longer-term data by averaging. In the case of monthly to yearly transformation, the averaging was carried out either across the whole year, or across subsets of the year (each month, or each quarter) as chosen by the expert. As with the response variable, many explanatory variables required transformation—often to their anomaly values (the anomaly of an observation with respect to a long-term mean). Other derived data (e.g., the distances of an observation to the ice pack edge and the nearest coast) were also included where appropriate.

Once the list of potential explanatory variables was finalized, the regression models were applied. Cross-validation (Stone, 1974) was used to select models and to assess their predictive accuracies. The process of list construction and model building was often iterated several times (indicated by the feedback paths in Fig. 3) before a final model was produced.

Results

We have chosen two examples to illustrate the data mining process. The first is a fairly well understood phenomenon, and we present the example as an intuitive problem against which we can explore and evaluate the discovery process. The second is a more speculative analysis. Each step in the process is marked by the corresponding label in Figure 3.

INDICATOR 56—MONTHLY FUEL USAGE OF THE ELECTRICAL GENERATOR SETS AND BOILERS AT AUSTRALIAN ANTARCTIC STATIONS

Special Antarctic blend, which is a light, diesel-like fuel, is used to power the electrical generators and boilers in each of the four Australian Antarctic stations. The fuel usage is measured from consumption gauges on each machine set and the total usage reported monthly. Fuel used elsewhere at the station, such as in the incinerators, vehicles, and water melt bells, is not included in this total.

The combined monthly fuel usage of the generator sets and boilers for Davis station (68°35′S, 77°58′E) from 1996 to 2001 was analyzed here. This fuel usage represents the fuel needed for both heating and powering the station. The heat generated by the electrical generators is used as the primary heat source for heating the station. During summer, this heat is often sufficient (or even excess to requirements), and the boilers are generally not used. During winter the boilers are used to provide additional heat to maintain the station temperature.

Problem Specification

We wished to determine which variables were the best predictors of station fuel usage. In fact, this is a relatively well understood problem: the fuel needed to heat each station is known to be dependent on the outside air temperature and wind conditions. Wind disturbs the layer of warm air that would otherwise envelope the building, increasing heat loss in a similar manner to the wind chill effect on the human body. The heating fuel needs also depend on various structural parameters such as the building volume and insulation efficiency. Our interest was in the intra-annual variations in fuel usage rather than the long-term (inter-annual) variations. The latter are largely determined by factors such as changes to the physical infrastructure of the stations. These changes were generally made during the summer months (between December and February). The monthly fuel usages were therefore normalized by subtracting the average fuel usage for each year.

Data and Preprocessing

The available explanatory variables comprised surface air temperatures (mean, lowest, and highest), mean lower stratospheric temperatures, mean mid-tropospheric temperatures, mean atmospheric pressure, electricity usage, mean wind speed, and the number of people on station (all measured at Davis station), and the sea surface temperature, sea surface temperature anomaly (the anomaly with respect to the long-term monthly average), and sea ice cover (measured adjacent to Davis). All explanatory variables were monthly. For the

TABLE 2

The relative importance of each variable in explaining the monthly fuel usage of the generator sets and boilers at Davis station (regression tree model). The importance value reflects the contribution that each variable makes as a splitting variable in the tree (maximum value arbitrarily scaled to 100%).

Explanatory variable	Importance (%)
Monthly mean air temperature	100
Highest monthly air temperature	89.8
Lowest monthly air temperature	89.2
Sea ice cover adjacent to Davis	80.2
Sea surface temperature adjacent to Davis	77.6
Monthly mean mid tropospheric temperature	77.0
Monthly mean lower stratospheric temperature	63.9
Sea surface temperature anomaly adjacent to Davis	25.6
Monthly mean of three-hourly wind speeds	19.1
Monthly electricity usage	15.4
Monthly mean atmospheric pressure	14.6

purposes of the demonstration, all of these variables were included as potential explanatory variables (there was no removal of irrelevant variables by the discipline expert).

Variable Selection and Model Building

Regression tree analysis yielded a tree with a single split. Months with a mean air temperature of less than –5.2°C had a higher fuel usage (an increase of 3200 L above the yearly baseline, $n = 56$), whereas warmer months had a lower fuel usage (a decrease of 6390 L below the baseline, $n = 22$). This simplistic model was relatively inaccurate (mean squared cross-validation error of 9.7×10^6 L^2, equivalent to an absolute error of 4.9% of the average monthly fuel usage). The relative importance of each explanatory variable in the regression tree is shown in Table 2. Surface air temperatures were the most important, followed by other variables with strong seasonal variations (e.g., sea ice cover and sea surface temperatures).

MARS analysis offered a more complex model in which fuel usage was determined by electricity usage, mean air temperature, and wind speed (Fig. 4). The model error was 4.2×10^6 L^2 (3.2%), less than half of that of the regression tree. The modeled effects of air temperature, wind, and electricity usage on fuel usage can be observed from Figures 4b and 4c. Colder air temperatures increased fuel usage, as did higher wind speeds and higher electricity usages. Both electricity usage and wind speed showed a threshold effect: increases of wind speed over 6.5 m s^{-1}, or electricity usage above about 160 MWh did not cause further increase in fuel consumption.

Evaluation against Domain Knowledge

These results are in good agreement with the known physical processes. However, for management purposes it would be preferable if electricity usage was not needed as an input. With this variable removed from the explanatory variable set, the MARS model used mean air temperature, wind speed, and mean lower stratospheric temperature as explanatory variables (Fig. 5). The model error (MSE 4.8×10^6 L^2, 3.5%) was slightly worse than that with electricity included. The model was particularly inaccurate during the summer months (note the truncation of the model estimates at low values in Fig. 5a). This is not surprising: as noted above, the summer fuel usage is essentially that needed for electricity generation. This result suggests that the electricity demand at Davis station is not well predicted by the remaining variables (which include station population). That electricity demand is independent of station population might seem to be counterintuitive; however, the majority of the electricity needs of the station are largely independent of personnel numbers (e.g., kitchen appliances, potable water management, hydroponics, refrigeration, and communications). Other needs such as lighting and general appliances tend to balance out: more lighting is needed in winter but there are fewer station personnel using appliances (J. Bonnice, unpublished data). The inclusion of lower stratospheric temperature as a explanatory variable was an indirect consequence of the fact that electricity usage during spring was lower than that during autumn. Surface air temperatures are roughly equal during these two seasons. However, lower stratospheric

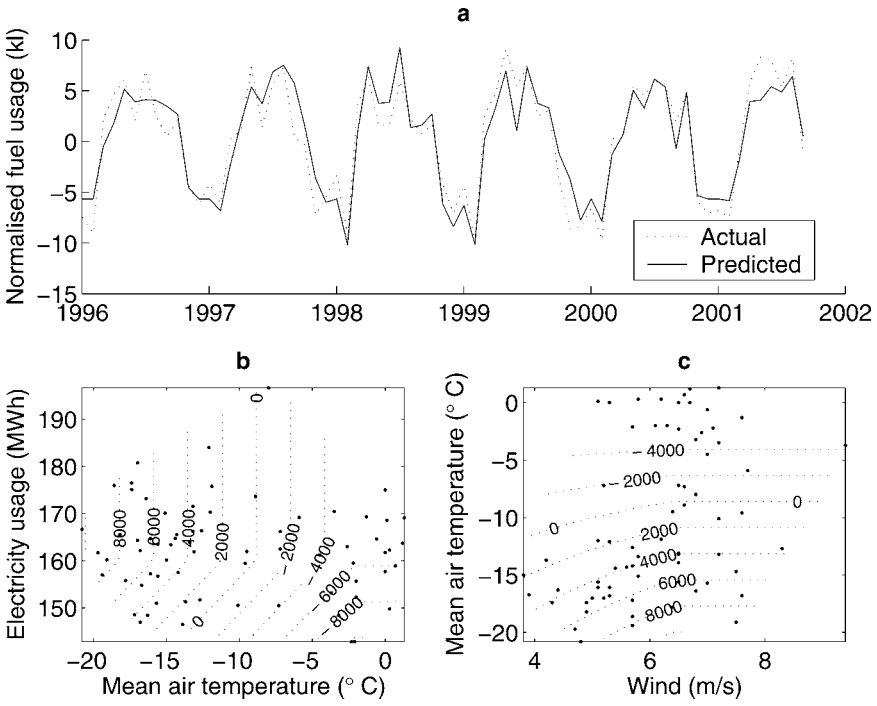

FIGURE 4. Multivariate regression spline model of fuel usage at Davis station. (a) Actual and predicted monthly fuel usages (fuel usages are shown as deviations from annual means, see text); (b) partial effects of air temperature and electricity usage on fuel usage; (c) partial effects of wind and air temperature on fuel usage. Contour lines (dotted) show predicted fuel usage, points show observed data. Predicted fuel usage increases with increased electricity usage or wind, or decreased air temperature.

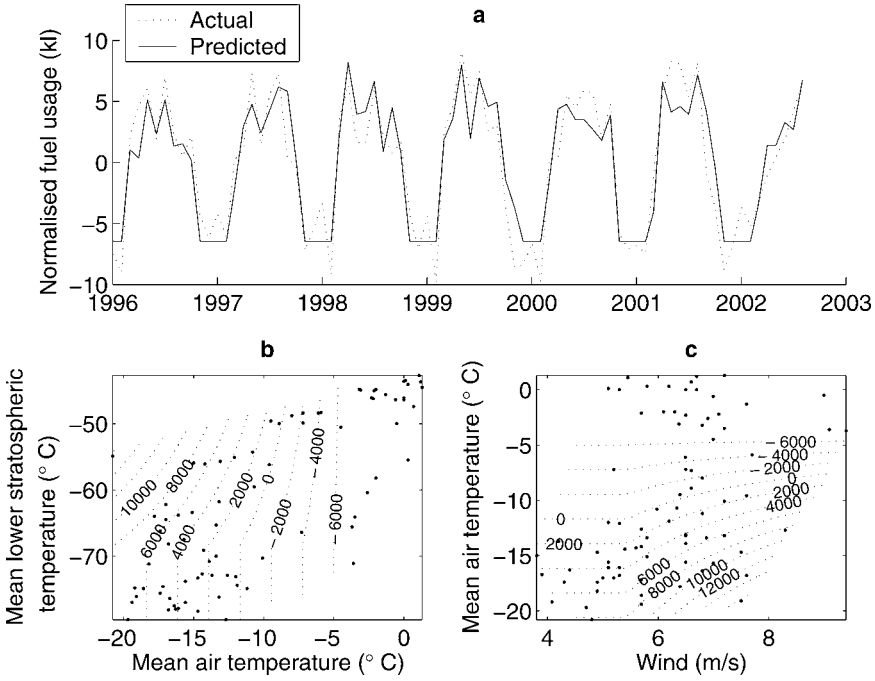

FIGURE 5. Multivariate regression spline model of fuel usage at Davis station, excluding electricity usage as an explanatory variable. (a) Actual and predicted monthly fuel usages (fuel usages are shown as deviations from annual means, see text). Note the truncation of the predicted values during the summer months when electricity demand dominates fuel usage; (b) partial effects of air temperature and lower stratospheric temperature on fuel usage; (c) partial effects of wind and air temperature on fuel usage. Contour lines (dotted) show predicted fuel usage, points show observed data.

temperatures lag the surface air temperatures by about one month. This allows the model to distinguish spring from autumn and adjust the fuel usage estimates accordingly. Excluding electricity usage did not affect the regression tree.

This relatively simple example demonstrates that both the tree and MARS methods are able to identify relevant explanatory variables from those available. The model generated by the MARS method was more representative of the true processes involved. The exclusion of electricity demand (which has a direct physical link with fuel usage) from the explanatory variable set yielded only a slight loss of predictive accuracy, but the model included a variable with no direct connection to the response variable. Inclusions such as this can be the result of a chance correlation (particularly if the number of data is few). Alternatively, as was the case here, the explanatory variable might be included to compensate for the lack of a more direct one. The two scenarios lead to quite different conclusions regarding the role of that variable. Differentiating between the two cannot be done numerically, and requires expert validation of the results in order to interpret the meaning of a variable in a given context.

INDICATOR 31—ANNUAL POPULATION ESTIMATES OF SOUTHERN ELEPHANT AND LEOPARD SEALS AT MACQUARIE ISLAND

Problem Specification

Observations of leopard seals (*Hydrurga leptonyx*) on Macquarie Island (54°30′S, 158°57′E) have been recorded since 1949 (note that this is a non-public indicator and does not appear in Table 1). While mature leopard seals normally reside in or near the outer edge of the Antarctic ice pack (Gilbert and Erickson, 1977), large numbers of juvenile leopard seals periodically occur on Macquarie Island, approximately every three to five years (Rounsevell and Eberhard, 1980; Rounsevell, 1988). No relationship between these periodic seal sightings and physical or biological environmental variables has to our knowledge been published. Ledingham (1979) investigated but did not find a relationship between variations in the proximity of pack ice and leopard seal abundance on Macquarie Island. A shortage of resources in the ice pack has been suggested as the most likely explanation for these periodic events (Rounsevell and Eberhard, 1980). Juvenile seals, being less adept at foraging in pack ice, might be out-competed for food and so forage northwards to subantarctic islands. The intent of this exercise was to use the data mining process to find relevant explanatory variables for these periodic leopard seal occurrences.

Data and Preprocessing

The seal abundance data (see Fig. 6) were collated from biology log-books, previous publications, and personal records of expeditioners. It is known that not all seals that visit the island are counted, and some (particularly during early years) may have been counted more than once. The accuracies of the raw abundance data were therefore variable. Further, the abundance data show evidence of a long-term decrease. These effects serve to obscure the periodic variations in the data, and so the abundance data were log-transformed and detrended. Those years during which large numbers of leopard seals were sighted (more than 2.5 times the fitted linear trend, $n = 19$) were marked as "leopard seal years" (see Fig. 6). Models were assessed on their ability to predict whether or not each year was a leopard seal year, thus making the problem one of classification rather than regression.

Few explanatory data were available for the period spanned by the seal observations. Those available comprised the lowest, highest, and mean monthly air temperatures, and mean monthly air pressure and wind speed at Macquarie Island (Shepherd, 2001). Mean monthly sea surface temperatures (in a $2° \times 2°$ area just south of Macquarie Island) were extracted from a global data set (Smith and Reynolds, 2003). Pack ice conditions are thought to be likely to contribute to this phenomenon (Rounsevell and Eberhard, 1980; Rounsevell, 1988). However, no reliable sea ice data were available for this period. Instead, mean monthly sea surface temperatures in waters just north of the estimated ice pack edge, immediately south of Macquarie Island, were used. The average monthly pack ice extent was calculated from satellite-derived estimates of sea ice cover from October 1978 to December 2001 (Cavalieri et al., 1999). All explanatory variables were calculated as yearly averages. We assumed that there might be a time

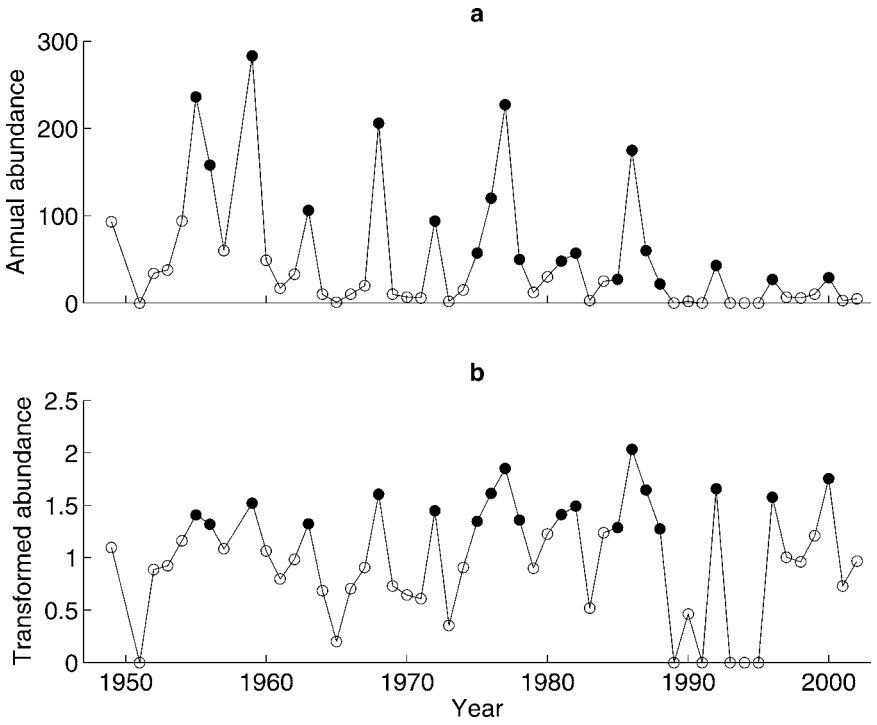

FIGURE 6. Annual abundance of leopard seals on Macquarie Island. (a) Raw data; (b) after log-transformation and linear detrending. Filled circles denote "leopard seal years": periodic occurrences of large numbers of leopard seals on the island.

lag between environmental conditions and leopard seal movements; therefore each explanatory variable was included in both immediate form (i.e., with the explanatory variable value taken from the same year as a seal population observation) and with a one-year lag (i.e., the explanatory variable value from the year prior to a seal population observation).

Variable Selection and Model Building

Classification tree analysis yielded a tree with a single split. In 14 separate years, Macquarie Island had an average highest monthly temperature of more than 8.9°C. On 11 of these 14 occasions, the following year was a leopard seal year. The cross-validated misclassification rate was 25%. This misclassification rate was lower than the null error rate of 37% (that obtained by simply guessing every year to be a non-leopard seal year). The next-best predictors were all warm temperature anomalies: warm mean monthly air temperature (>4.8°C) in the previous year, warm sea surface temperature (>6.7°C) adjacent to Macquarie Island in the previous year, and warm lowest monthly air temperature (>–0.5°C) in the previous year. Of the seven explanatory variables that yielded better classification accuracy than the null, six were from the year previous to that being classified. Adding further splits to the tree led to a decrease in predictive accuracy, indicating overfitting of the tree to the data.

Evaluation against Domain Knowledge

Interpretation of the analyses at once suggests that the Antarctic circumpolar wave (ACW; White and Peterson, 1996) has a role in leopard seal haul-outs at Macquarie Island. The circumpolar wave is known to produce two suites of contrasting events at four to five year intervals. One suite of characteristics in the ocean and atmosphere south of Macquarie Island is a warmer sea surface temperature, less extensive sea ice northwards, higher sea levels, and poleward wind anomalies. The alternate suite reverses these anomalies. From the results of the classification analyses it appears that leopard seals tend to move northwards to Macquarie Island during the warmer sea phase. It is likely that it is not the sea surface temperatures that drive this behavior directly, but that these measurements act as a proxy for more direct factors, such as characteristics of the sea ice or prey abundance. The results indicate that variables from the previous year were better predictors than the same variables during the leopard seal year itself. This suggests that there might be a time lag between environmental conditions and the response of the leopard seals, possibly reflecting intermediate steps such as the response of prey species to environment. Alternatively, the seals might be responding rapidly (i.e., with a time lag of less than one year) but to environmental conditions to the east of Macquarie Island. The ACW propagates eastward (White and Peterson, 1996), and so the warm phase of the ACW to which the seals might be responding would have been recorded at Macquarie Island during the previous year.

Problem Iteration

With these initial results in hand, we revisited the problem, this time focusing on the spatial relationships of the data, in particular the sea surface temperature and sea ice variables. We explored correlation maps of leopard seal abundances with gridded sea surface temperature data (available for all years) and gridded sea ice concentration data (available only from 1979 onwards). Figure 7 shows these maps for sea ice and sea surface temperature in the previous year. Leopard seal abundances were negatively correlated with sea ice cover in the region south and slightly to the east of Macquarie Island in the previous year (Spearman $r = -0.65$, $p < 0.05$; Fig. 7a). Areas of positive correlation at approximately 120°E and 140°W can also be seen on this figure; these correspond to the cold (more extensive sea ice) nodes of the ACW. Leopard seal abundances were also correlated with warm sea surface temperatures in the region of Macquarie Island in the previous year (Fig. 7b). These correlations were weaker and did not reach statistical significance, suggesting that the link between leopard seal behavior and sea ice conditions might be more direct than that with sea surface temperatures. The correlations for sea ice and sea surface temperature conditions of the same year (not shown) were generally weaker than those of the previous year. More data about the role of sea ice and associated ecosystems at different latitudes in the lives of leopard seals are required to take explanations further. However, it is

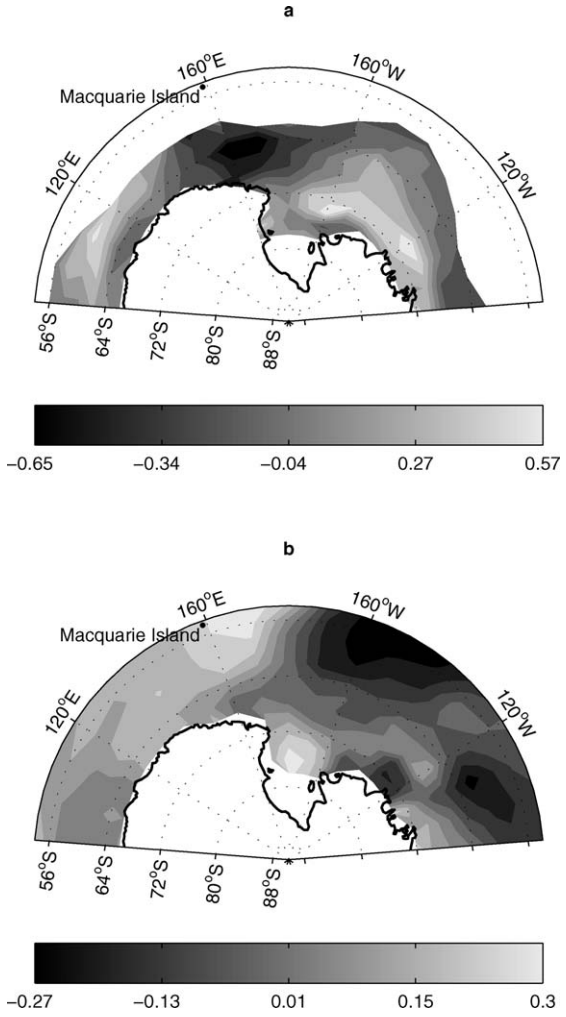

FIGURE 7. Correlation maps of annual abundance of leopard seals on Macquarie Island with (a) average sea ice concentration from the previous year and (b) average sea surface temperature from the previous year. Leopard seal abundances are correlated with low sea ice cover south of Macquarie Island, and more weakly with warm sea surface temperatures in the Macquarie Island region.

clear that short-term climate periodicities at sea south of Macquarie Island have importance in determining northwards movements of leopard seals. These results will contribute to ongoing research in this area (Burton, 1998–2003).

Discussion

The examples presented here demonstrate that a data mining approach can be used to identify functional relationships within a collection of data. The approach has obvious application to developing and refining scientific models. It also offers promise as a unique search tool to assist scientists in navigating the holdings of a data center, by searching for data that is functionally related to a data set of interest.

Our experiences with Antarctic scientific data have identified some particular challenges for data mining. Antarctic data sets tend to be sparse, containing only a few observations. Often a data set will be collected for a specific project and span only one or two seasons. Antarctic data are also diverse, covering many different types of data, sampling frequencies, and acquisition methods. For the model selection problem, this means that there are often many potential explanatory variables but few observations on which to base the selection algorithms. This is in contrast to typical applications of data mining in business and other scientific fields, in which data volumes are often large and homogeneous. Further, the explanatory data available for Antarctic data are often not directly relevant to the response variable. This was demonstrated with the leopard seal data. The phenomenon of leopard seals on Macquarie Island has previously been postulated to be due to food scarcity in the ice pack. However, none of the available explanatory variables were direct measures of prey abundance. In these situations, the resultant models will tend to have weak predictive power. A similar problem can arise from the spatial sparsity of the data. Antarctic data tend to be very patchy in their spatial distributions. Analyses of such data therefore lead to local conclusions, which can not necessarily be extended to wider regions of the Antarctic.

These difficulties are not easily overcome. With small data sets, methods that make limited use of the information provided by data can perform poorly. The regression tree model of fuel usage presented here demonstrates this. In such cases, the prior knowledge of the experts becomes an important source of information. This information was incorporated here in an indirect manner, by allowing the analyst to constrain elements of the search procedure and to manipulate variables. Other techniques, such as Bayesian networks (e.g., Heckerman, 1999) incorporate prior knowledge in more formal manner. Such models could be used as replacements for the regression techniques included here. However, during the early stages of model development, there is often little prior knowledge available. In these instances, identifying only a single model to explain the observed data is a poor strategy. A better approach is to identify a number of likely models. These may have contradictory interpretations in terms of physical processes. Further data collection or experimentation, coupled with expert interpretation, could then resolve the ambiguities. Careful design of these subsequent experiments can reduce the amount of extra data needed (e.g., Ramakrishnan and Bailey-Kellog, 2002). Thus, the iterative nature of the data mining process can extend across the entire cycle of scientific investigation (Hand, 1999; Ramakrishnan and Grama, 2001).

Even when data are plentiful, identifying only one candidate model can be a risky proposition. Often, several of the explanatory variables yield similar model accuracy. Investigating only one possible choice risks the drawing of limited or incorrect inferences. This can be particularly problematic with highly correlated predictors, which is a common occurrence in environmental databases such as the SIMR used here. In these analyses we have typically examined a range of possible models, drawing on the knowledge of the relevant expert to evaluate each. The ranked list of explanatory variable importances at each split (trees) or knot placement (MARS) was also examined to identify a number of possible variables at each stage of a model.

The reuse of archived data in the manner described here must be undertaken with care. It is the responsibility of the analyst to ascertain the suitability of data for reuse, including examining the methods used to collect the data in the first instance. This is often a problem with observational data (e.g., wildlife sightings) collected opportunistically and without a balanced sampling strategy (Raymond and Woehler, 2003). In some instances, databases of such sightings can be resampled to approximate a more rigorous experimental design (Guisan and Zimmermann, 2000). If the database is large, the subsequent reduction in data volume might also be beneficial. Other aspects of the model must also be considered. For example, a large proportion of the data held by the AADC have spatial and temporal components. Ignoring spatial correlation can lead to erroneous conclusions (Carroll and Pearson, 2000; Keitt et al., 2002). Many techniques for mining spatial and temporal data have been developed—see, for example, Koperski and Han (1995), Ng and Han (2002), and the bibliography provided by Roddick and Spiliopoulou (1999). In the leopard seal example we used

spatial correlation maps to explore the relationships between data sets. For large spatial databases, a more computationally efficient approach may be required (e.g., Zhang et al., 2003).

Noise or uncertainty in data can be difficult to deal with. Remote sensed data can have large uncertainties. For example, estimates of sea ice cover derived from satellite images are known to be poor when confronted with newly forming sea ice (Cavalieri et al., 1999). Long-term Antarctic data sets have almost without exception been collected by a variety of researchers or instruments, giving uncertainties that vary across the history of the data. However, incorporating uncertainty information into analyses can provide important guidance for future experiments by identifying areas where the support for a hypothesis or the understanding of causal mechanisms is weak. The AADC encourages Antarctic scientists to describe the likely uncertainties in their data in the associated metadata records; these uncertainties are then included as fields in the relevant databases. This information could be used, for example, in a Monte Carlo approach to uncertainty analysis. This would examine the effects on a model of random perturbations to the explanatory data within the limits of these uncertainties. Regions of explanatory space in which a model is particularly susceptible to noise might indicate a weakness of the model, or that further experimentation could profitably be directed to this aspect of the model's behavior.

Acknowledgments

Ian Ball, Leon Barmuta, Lee Belbin, Glenn De'ath, and two anonymous reviewers provided critical reviews of earlier versions of this manuscript.

References Cited

Agrawal, R., Imielinski, T., and Swami, A., 1993: Mining associations between sets of items in massive databases. *In:* Halevy, A. Y., Ives, Z. G., and Doan, A. (eds.), *Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data*. Washington D.C., 207–216.

Apte, C., Grossman, E., Pednault, E., Rosen, B., Tipu, F., and White, B., 1999: Probabilistic estimation-based data mining for discovering insurance risks. *IEEE Intelligent Systems,* 14: 49–58.

Belbin, L., Gibson, J., Davis, C., Watts, D., and McIvor, E., 2003: State of the Environment reporting: an Antarctic case study. *Polar Record,* 39: 193–201.

Bergstrom, D., and Chown, S., 1999: Life at the front: history, ecology and change on Southern Ocean islands. *Trends in Ecology and Evolution,* 4: 472–477.

Breiman, L., Friedman, J., Olshen, R., and Stone, C., 1984: *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Burton, H., 1998–2003: Periodicity in foraging areas of leopard seals in the region south of Macquarie Island. Australian Antarctic Science Project 2268. http://aadc-maps.aad.gov.au/aadc/metadata/metadata_redirect.cfm?md=AMD/AU/ASAC_2268. Site last accessed 20 May 2005.

Carroll, S., and Pearson, D., 2000: Detecting and modeling spatial and temporal dependence in conservation biology. *Conservation Biology,* 14: 1893–1897.

Cavalieri, D., Parkinson, C., Gloerson, P., and Zwally, H., 1999 (updated 2002): Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data. Boulder, CO: National Snow and Ice Data Center, CD-ROM.

Chaudhuri, S., 1998: Data mining and database systems: Where is the intersection? *IEEE Data Engineering Bulletin,* 21: 4–8.

Crawford, J., and Crawford, F., 1996: Data mining in a scientific environment. *In:* Bossomaier, T., and Chubb, L. (eds.), *Proceedings of AUUG 96 & Asia Pacific World Wide Web 2nd Joint Conference*. Melbourne, Australia.

Friedman, J., 1991: Multivariate adaptative regression splines. *Annals of Statistics,* 19: 1–141.

Friedman, J. H., 1997: Data mining and statistics: What's the connection? *In:* Scott, D. (ed.), *Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics*. Mining and Modeling Massive Data Sets in Science, Engineering, and Business with a Subtheme in Environmental Statistics, Houston, TX, 3–9.

Gilbert, J., and Erickson, A., 1977: Distribution and abundance of seals in the pack ice of the Pacific sector of the Southern Ocean. *In:* Llano, G. (ed.), *Adaptations within Antarctic Ecosystems*, Washington, D.C.: Smithsonian Institute, 703–740.

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P., 1996: Statistical inference and data mining. *Communications of the ACM,* 39: 35–41.

Guisan, A., and Zimmermann, N., 2000: Predictive habitat distribution models in ecology. *Ecological Modelling,* 135: 147–186.

Hand, D., 1999: Statistics and data mining: Intersecting disciplines. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery & Data Mining,* 1: 16–19.

Hastie, T., Tibshirani, R., and Buja, A., 1994: Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association,* 89: 1255–1270.

Hastie, T., Tibshirani, R., and Friedman, J., 2001: *The Elements of Statistical Learning*. New York: Springer.

Heckerman, D., 1999: A tutorial on learning with Bayesian networks. *In:* Jordan, M. (ed.), *Learning in Graphical Models*. Cambridge, MA: MIT Press, 301–354.

Keitt, T., Björnstad, O., Dixon, P., and Citron-Pousty, S., 2002: Accounting for spatial pattern when modeling organism-environment interactions. *Ecography,* 25: 616–625.

Koperski, K., and Han, J., 1995: Discovery of spatial association rules in geographic information databases. *In:* Egenhofer, M. J., and Herring, J. R. (eds.), *Proceedings of the 4th International Symposium on Advances in Spatial Databases*. London: Springer-Verlag, 47–66.

Ledingham, R., 1979: *The biology of the leopard seal* Hydrurga leptonyx *(de Blainville) with special reference to Macquarie Island*. Ph.D. thesis, Scott Polar Research Institute, Cambridge, England.

Mannila, H., 2000: Theoretical frameworks for data mining. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery & Data Mining,* 1: 30–32.

Mannila, H., Toivonen, H., and Verkamo, A. I., 1997: Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery,* 1: 259–289.

Miller, A., 2002: *Subset selection in regression*. Second edition. Boca Raton, Fl: Chapman & Hall/CRC.

Ng, M., Huang, Z., and Hegland, M., 1998: Data-mining massive time series astronomical data sets—a case study. *In:* Wu, X., Ramamohanarao, K., and Korb, K. B. (eds.), *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery in Data Bases*. Melbourne, Australia, 401–402.

Ng, R., and Han, J., 2002: CLARANS: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering,* 14: 1003–1016.

Potter, C., Tan, P.-N., Steinbach, M., Klooster, S., Kumar, V., Myneni, R., and Genovese, V., 2003: Major disturbance events in terrestrial ecosystems detected using global satellite data sets. *Global Change Biology,* 9: 1005–1021.

Ramakrishnan, N., and Bailey-Kellog, C., 2002: Sampling strategies for mining in data-scarce domains. *IEEE/AIP Computing in Science and Engineering,* 4: 31–43.

Ramakrishnan, N., and Grama, A., 2001: Mining scientific data. *Advances in Computers,* 55: 119–169.

Raymond, B., and Woehler, E., 2003: Predicting seabirds at sea in the Southern Indian Ocean. *Marine Ecology Progress Series,* 263: 275–285.

Reynolds, R., Rayner, N., Smith, T., Stokes, D., and Wang, W., 2001:

An improved in-situ and satellite SST analysis for climate. *Journal of Climate,* 15: 1609–1625.

Rocke, D., and Dai, J., 2003: Sampling and subsampling for cluster analysis in data mining: with applications to sky survey data. *Data Mining and Knowledge Discovery,* 7: 215–232.

Roddick, J., and Spiliopoulou, M., 1999: A bibliography of temporal, spatial, and spatio-temporal data mining research. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery & Data Mining,* 1: 34–38 (updated version available at: http://kdm.first.flinders.edu.au/IDM/STDMBib.html). Site last accessed 20 May 2005.

Rounsevell, D., 1988: Periodic irruptions of itinerant leopard seals within the Australasian sector of the Southern Ocean, 1976–86. *Papers and Proceedings of the Royal Society of Tasmania,* 122: 189–191.

Rounsevell, D., and Eberhard, I., 1980: Leopard seals, *Hydrurga leptonyx* (Pinnipedia), at Macquarie Island from 1949 to 1979. *Australian Wildlife Research,* 7: 403–415.

Schwabacher, M., and Langley, P., 2001: Discovering communicable scientific knowledge from spatio-temporal data. *In:* Brodley, C. E., and Danyluk, A. P. (eds.), *Proceedings of the Eighteenth International Conference on Machine Learning.* Williamstown, MA, 489–496.

Shepherd, D., 2001: Meteorology data from Macquarie Island Station (300004), 1948 on-going, surface measurements. Australian Antarctic Data Centre—SnoWhite Metadata (http://aadc-maps.aad.gov.au/aadc/metadata/metadata_redirect.cfm?md=AMD/AU/Macquarie_met). Site last accessed 20 May 2005.

Smith, T., and Reynolds, R., 2003: Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *Journal of Climate,* 16: 1495–1510.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions (with discussion). *Biometrika,* 64: 29–35.

White, W., and Peterson, R., 1996: An Antarctic circumpolar wave in surface pressure, wind, temperature and sea ice extent. *Nature,* 380: 699–702.

Zhang, P., Huang, Y., Shekhar, S., and Kumar, V., 2003: Correlation analysis of spatial time series datasets: A filter-and-refine approach. *In:* Whang, K.-Y., Jeon, J., Shim, K., and Srivastava, J. (eds.), *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '03),* Lecture Notes in Artificial Intelligence Volume 2637, pp. 532–544.

Revised ms submitted October 2004