Data Mining for Ecological Field Research: Lessons Learned from Amphibian and Reptile Activity Analysis

Jeffrey D. Campbell
Center for Urban Environmental Research and Education, UMBC
jcampbel@umbc.edu

Karyn Molines, Christopher W. Swarth
Jug Bay Wetlands Sanctuary
RPKARYN@aacounty.org,
RPSWAR96@aacounty.org

Abstract

While there are many similarities to traditional data mining, analysis of ecological field research data presents some uncommon data and algorithmic requirements. Data challenges reflect the broad range of data that is relevant to analysis in this domain. Much of that data is obtained from physical measurements that are subject to mechanical, electrical or human error factors. For some types of data, for example meteorological observations, multiple sources are publicly available. One challenge is to properly distinguish errors from microclimatic variations which are important to ecologists. Varied units of measure and inconsistent time policies are also commonly found in environmental data.

Requirements for data mining algorithms for animal activity analysis include computing differential and integral values, handling highly interrelated environmental factors, and strong support of time series analysis. This includes both variable types of lag times and more types of cyclical effects than usually found in business data.

1. Introduction

This paper describes the data and analysis requirements for data mining investigating the effect of various environmental factors on animal behavior. Three motivating examples are provided, but the focus is on identifying more generalized requirements from this case study.

A data warehouse was developed for Jug Bay Wetlands Sanctuary (JBWS) in Maryland, on the Patuxent River which is a tributary to the Chesapeake Bay. From an ecologist's perspective the goal was to provide a data repository for meteorological and related data that was relevant to the broad spectrum of research projects at the site including salamander migration, turtle movement and nesting, other amphibian activity, vernal pool conditions, terrestrial and submerged aquatic vegetation growth and bird migration.

From a data mining research perspective, the warehouse was the first step in developing and testing new techniques particularly appropriate for such physical and biological data.

The remainder of the introduction describes the application domain. Section 2 focuses on the heterogeneous data and the related quality, data cleansing and metadata problems that were frequently encountered in the environmental data. Section 3 describes the requirements for data mining algorithms to specifically address ecological research questions.

1.1. Benefits of Ecological Data Mining

In addition to the inherent increase in knowledge from an improved understanding of amphibian and reptile activity, there are other benefits to the potential results from improved data mining.

Amphibians are sensitive to the environment and are often monitored as a measure of change in conditions [1]. For example, the FrogWatch USA program [2], among others, collects information from volunteers about which species of frogs and toads are calling during mating season. The frequency at which the data is reported is highly variable. Analysis of the best conditions for observation could make the data collection more efficient. It could also identify changes resulting from selecting better or poorer observation times instead of real changes in the amphibian population. Observation time selection may not be random if volunteers tend to avoid unpleasant weather conditions that are later found to be prime conditions for the animals being observed.

1.2. Motivating Examples

Two sets of monitoring data were selected for the initial analysis. The first consists of data on marbled salamanders captured during their annual breeding migration. The relevant data is the number of individuals caught, the specific locations, the date and limited information on the direction of movement (toward or away from the breeding area). Over 10,000 salamanders have been caught and released since the study began in 1988.

The second dataset is the movement of box turtles that were monitored with radio telemetry. Periodically researchers would find each of the turtles, record the location and compute the net distance traveled.

The data warehouse structure proved to be reusable and was also populated with weather data from another county to be used to analyze toad and frog calling during mating season. This dataset consisted of observations made by 85 volunteers at 46 locations in one county in the last four breeding seasons as part of FrogWatch USA [2]. The relevant data is the location, date, time, temperature, which species were heard, the activity level (ordinal scale of 0 to 3) and some general weather observations. There were over 1500 site visits with over 2700 species-day-location observations.

Since reptiles and amphibians control their body temperature through external methods, weather factors would seem to be important. Rainfall was found to be particularly important and proved to be most difficult to interpret [3]. Without rain, very few if any salamanders are caught. When it rains, the number caught can vary by two orders of magnitude with no obvious relationship to the amount of rain (Figure 1). Seeking an explanation of this variability was the motivating factor in using data mining.

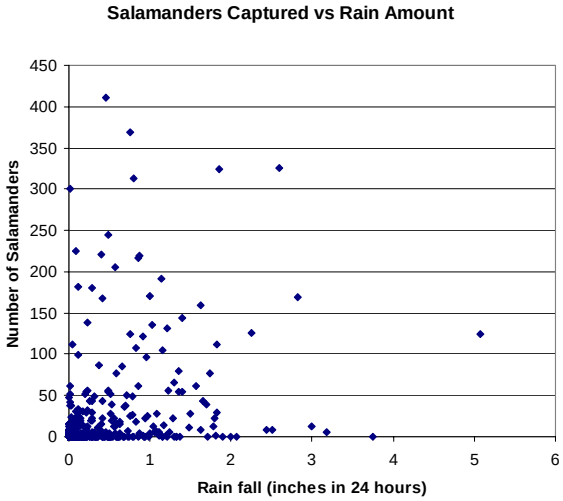

Figure 1. Daily salamander capture and precipitation

2. Data Requirements

This section first summarizes the types of data relevant to the research in order to provide context. The next subsection addresses some data quality and metadata issues and that are typical in this domain. Weather data presents an opportunity and challenge in the area of data cleansing. The last subsection describes an interesting biological nuance to the typical missing data and zero value problems.

2.1. Data Overview

Relevant weather data includes precipitation, humidity, wind speed and direction, temperature and barometric pressure. Daily weather temperature extremes and precipitation have been recorded manually from a weather station in an open meadow at JBWS since 1988. Internet data is increasingly available from the National Weather Service (NWS), networks of amateur observers posting data to websites such as the Weather Underground [4], CWOP [5] and proprietary collections, e.g. Weather Bug [6]. Data was obtained from a NOAA National Weather Service weather station 15 km from the site with hourly weather data starting in 1943. Since mid-2003, an automated station located a marsh at the sanctuary has recorded weather conditions every 15 minutes as part of NOAA's National Estuarine Research Reserve System. Late in 2006, an observer unaffiliated with the sanctuary started posting weather data to the Internet apparently from a backyard location 3 km from the sanctuary

Solar and lunar affects on lighting could be important. The time for sunrise and sunset is obvious since that determines the number of hours of light (or dark depending upon the research emphasis). Some plant and animal activity has been associated with the length of the day. The daily phase of the moon, percent of moon being illuminated and moonrise and moonset times are also included. Combined with cloud cover data, this would provide an approximate amount of light at night.

Solstice and equinox dates were included for computation of days within an astronomical season. These would also be correlated to the angle of sun elevation at noon and the compass direction of apparent sun rise and sun set. Both of those could determine the amount of sun in a partially shaded area.

The timing and magnitude of tides have a clear impact on the environmental conditions. Both values vary geographically. For example, in the tidal river at JBWS along a 5.2 km distance along the river there is a 49 minute difference in predicted high tide time and 51 minute difference in low tides. In tidal areas around the Chesapeake Bay, winds and other factors result in actual tide magnitudes that vary substantially from the predictions.

Characteristics such as the age, size and sex of the animals themselves may be significant. For the marbled salamanders, the males tend to migrate before the females [3] while smaller males arrive before larger males [7].

2.2. Data Quality and Metadata

Many of the typical data issues found in traditional data mining applications are also found in ecological data. Variable levels of metadata and quality assurance/ quality control were provided with data from different sources. Next, this section describes two pervasive data issues – units of measure and time. Other examples of environ-

mental metadata uncertainty are briefly mentioned. A problem with precipitation is described in detail to give a sense of the complexity of physical data measurement.

2.2.1 Quality Assurance/Quality Control One of the weather data sources provided a code indicating the quality of each value for each observation. The review process was apparently automated but full documentation was not provided. It appears that it flags values outside of two or three standard deviations from the mean. Unfortunately the statistics are based on the year, so correct values during summer heat waves are detected as outliers but an obviously high temperature in winter could be missed.

Automatically collected data and manually recorded data could exhibit different types of errors. A data mining approach could be taken to classify data points as valid or suspect. However, there are some quality problems that are difficult to detect. One FrogWatch volunteer did not follow the observation protocol to use a thermometer to take the temperature at the observation site but just called the local utility company's time and temperature number and reported that temperature as the observation.

2.2.1 Pervasive Issues Presumably since much of the weather data is intended for public information, the units of measure are commonly Fahrenheit and English units which are easily converted to metric. The decision was to store data in the warehouse in the observed units. Data views were defined with unit of measure conversions.

The time stamp for observations proved to be a more difficult issue to resolve in the data warehouse. Data was obtained with the following time policies:

- Universal Coordinated Time (UTC), formerly called Greenwich Mean Time
- Standard time all year
- Local time (standard time or daylight saving time depending upon the date)
- One source indicated that "current" data followed one policy but "archived" data followed another.
- Another source indicated that local time was used but examination of the data at the transition between standard time and daylight time showed no evidence of a change.
- Unspecified

To add to the complexity, rules setting the transition dates between standard and daylight saving time have changed this year and four other times since 1966 [8].

Due to the apparent uncertainty in some of the time policies, the decision was made to store the data in the original format. A metadata table was dedicated to time offsets. This allows views and queries to convert to the appropriate time standard. A field indicating the use of daylight saving time for each date was computed and added to the daily table that was the center of the star schema. Perhaps the data mining time series algorithms could be enhanced to handle the varied time policies more efficiently.

2.2.2 Metadata Uncertainty. There is room for interpretation of some commonly reported weather values. Averages are sometimes reported without specifying the frequency of the data used in the computation. For example, averaging the minimum and maximum temperature to compute the mean would not necessarily give the same value if the temperature every five minutes was averaged. Average wind direction can be weighted by wind speed as a vector or simply the frequency of the direction. The use of magnetic or true north is rarely specified, but for areas of low declination, the coarseness of the direction measurement is generally larger than this uncertainty. The National Weather Services reports three values for barometric pressure with different definitions, but the other sources only report one value without defining it fully.

2.2.3 Precipitation The rain fall data from other Internet sources that aggregate data submitted by many people from different sites was more difficult to interpret. A fundamental source of uncertainty may be the common use of "tipping bucket" rain gauges. In these devices, a funnel directs rain water onto a small container that empties itself (by tipping) after a specified amount (often .01 inches) of rain accumulates. The physical movement of emptying of the container is recorded electrically. Based on a brief survey of the specifications for several such devices commonly used for the amateur Internet data, these devices are polled periodically (2-4 minutes) by the indoor base station. Data is uploaded to the website periodically from the base-station, typically at 5 or 15 minute intervals. Since the polling interval is a substantial fraction of the upload interval, the number of polling events per upload varies reducing precipitation timing accuracy. Furthermore, the polling interval for at least one rain gauge is different for wired and wireless operations. In this case, even knowing the model number of the device is not sufficient to determine the data frequency.

The second data problem is that the interpretation of the "Hourly Precipitation" field posted to the Internet is uncertain. Without clear metadata it could potentially be

- Rain in the last 60 minutes
- Rain since the beginning of the current hour (clock time)
- Rain for an hour based on the current rate. This computes the amount of rain per unit time in the last interval and extrapolates that to an hour. This reports rain intensity rather than accumulation.

Values for the first two could be reported frequently (typically every five minutes) or just once per hour with zero reported between the hourly values. If the former, accumulated rain must be calculated as a rolling difference, not simply the total of the data reported. Reporting the later more directly reports accumulation, but the time granularity is an hour.

The definition given in the protocol for uploading data to one of the weather data aggregation sites defines the

field as “the accumulated rainfall in the past 60 mins.” Unfortunately the data from the various stations using this web site are not consistent with this definition. The computation of rainfall for each period to determine the total gives numerous periods with negative computed rain. It appears that at least some stations are reporting values other than specified in the protocol.

Research continues to work to resolve these differences. Perhaps there is a data mining approach that could be used to classify the stations with respect to the true nature of the data being reported as “hourly rain.” This solution would appear to have more widespread applicability with another web aggregator of weather data simply providing a disclaimer that precipitation data is what was supplied by the observer and is not consistently reported.

2.3. Data Cleansing

Some weather values vary within short distances with different ecological niches resulting from differences in microclimate. For example, temperature, humidity and wind are likely to differ between a heavily shaded forest and an adjacent sun-baked meadow. Other measurements, such as barometric pressure are less likely to exhibit such short distance variations.

Manual observations at variable locations, such as the air temperature at Frog Watch sites are also subject to variation in microclimate. For example, one of the authors observed a difference of 8°F depending upon the exact location near the pond being observed. The observation protocol did not specify which value to report.

Beyond the obvious care in selecting data sources, if multiple sources are available, it is important for the data cleansing process to distinguish between errors and microclimate differences. It is tempting to use multiple sources to cross-validate the data but this could unintentionally remove the effect being sought.

2.4. Missing and Zero Data

In addition to the usual problems with missing and zero data, two examples were encountered that suggest the need for extra care in analysis. Ironically, during a recent 24 hour intensive biological survey neither of the two nearby automated data sources was recording data resulting in traditional “missing” data.

Temperature extremes and precipitation have been recorded manually at JBWS. However, data for unstaffed days is aggregated in the next reading. This means that data value for the unstaffed day(s) is null, but the data is not exactly “missing” since it is included in the next staffed day’s value. However, the data for that next staffed day is also not correct for that date.

Another type of partially missing data can occur. Occasionally a predator will be caught or a capture bucket will fill with water allowing some or all of the salamanders to escape. When such conditions are detected, it is noted in the data files. The number caught should be treated as a minimum, not the true value. Unfortunately, water overflows occur during rain storms which are when they would have been most likely to catch salamanders, so the data loss to flooding is not random.

There is another aspect of observing animals in the field that goes beyond the typical missing data problem. In essence, not observing an animal at a particular time does not mean it was not there. It may have been there and not seen/heard or it may have been absent.

The meaning of zero can also vary depending upon the type of analysis. The frog data provides a good example. The web-based data entry form lists all species found in the entire state of Maryland with a default value of zero calling activity. However, no site would have all species present. All of the zero observations are useful when the question is which species are found in which locations. However, if the analysis is which species call based on different weather conditions, the data indicating they were not calling at sites where they do not live is confounding data.

3. Requirements for Algorithms

Results from other salamander research illustrate the types of relationships that have been identified which. At a minimum, data mining algorithms should be able to identify those types of relationships. Ideally, the new algorithms would go beyond the known relationships and enhance the models. This section first describes this related work. Requirements for time series analysis including periodicity, differential and integral values and lag times are then identified. Finally, a caveat about interrelated data is presented.

For almost 100 years researchers have attempted to determine the environmental factors related to the annual breeding migration of salamanders. Maximum and minimum temperatures were not correlated with the arrival of spotted salamanders at breeding ponds. Instead, breeding activity was dependent on evening spring rainfalls, but only when temperatures were above freezing [9]. Soil temperature in addition to rainfall was correlated to spotted salamander migration [10]. They determined that threshold values of rainfall (4 mm) and mean air temperature (5.5 C) only explained variations of salamander activity when soil temperature at 30 cm below the surface was at least 4.5 C and the thermal profile was reversed (i.e., the surface was warmer than the soil at 30 cm). For mole salamanders, the cumulative rainfall during the breeding season was highly correlated to their yearly breeding population size [7]. This related ecological research was described to give a sense of the types of effects that have been found. However, the fall migration of the marbled salamanders could differ from the spring spotted and winter mole salamander activity.

Daily, weekly, monthly and annual cycles could easily occur in business data. There is no reason to expect there to be weekly cycles in natural phenomena. However, the lunar cycle could be important. High and low tide patterns are predictable from solar and lunar positions with factors for geography but the actual tides are also influenced by weather. Seasons, as measured by solar position can also be significant. There can also be interactions between all of these cyclical factors. There are also longer term population cycles. Weather conditions in the spring can substantially impact the number of larvae that survive. Years later when they reach reproductive age, the survival rate would be an important factor in predicting the number captured.

Instead of or in addition to the magnitude of the various parameters, the derivative or integral of the values could be significant. The rate of change of the value (first derivative with respect to time) or the rate of change of the rate of change (second derivative) could be significant. For example, falling barometric pressure before a low pressure system could predict rain with an increased rate of change as the low approaches. The cumulative value of the stimulus over time could be more important than the immediate value. The total number of days or hours above a temperature threshold has been found to predict blooming of plants in spring.

There could be time delays between the stimulus and response. A critical value, for example some amount of rain, might need to be reached before there was a response. Response could be proportional to the stimulus or non-linear. Perhaps a stimulus after a longer absence is more powerful, for example rain after a longer than usual dry period. The sequence of events could be important, for example, plant seeds not germinating until after a hard freeze.

There are numerous relationships between weather data elements that are either causal or correlated. Relative humidity is high when it is raining. Summer thunderstorms have rain fall, higher winds and a temperature decrease. The passage of a weather front is similarly detectable through various measures. These interrelationships may make it more difficult to distinguish the primary factor in behavior. This suggests that associative data mining techniques are likely to identify many relationships between factors that are independent of the animal activity of interest.

5. Conclusions and Future Work

During the development of the data repository and initial analysis a number of key requirements have been identified. Not unexpectedly, inconsistencies in data definitions and data quality issues are present. The varied sources of weather data provide both an opportunity for better data quality review but the challenge of not obscuring the small microclimatic differences that are important to the ecological research questions.

Time-based analysis is critical to much ecological research. There can be multiple cyclical effects. It is likely that there are time lags between environmental changes and behavioral responses. Responses depending upon the first or second derivative of the factor with respect to time or cumulative values may also exist. Analysis of weather and other factors between arbitrary dates and times for animal data is another aspect.

With many meteorological values interrelated to each other, identifying the specific factor for a model from field data is a challenge. Like all data mining, domain expertise from ecologists is critical in developing models with believable causality not just spurious correlations.

While the project started with a small scope focused on weather factors for salamander and turtle activity, the scope can easily expand. NWS data is available across the country. There are 27 areas in the NOAA National Estuarine Research Reserve system many with local weather and water quality data in the same format.

The current analysis addressed location questions primarily in the context of micro-climate. A richer geospatial representation would support more elaborate analysis and provide additional data mining requirements.

10. References

[1]U. S. EPA. 2002. *Methods for Evaluating Wetland Condition #12 Using Amphibians in Bioassessments of Wetlands*, Office of Water, U.S. Environmental Protection Agency, Washington, DC. EPA-822-R-02-022.

[2] FrogWatch USA, National Wildlife Federation and US Geological Survey, www.nwf.org/frogwatchusa

[3]Molines, K and Swarth, C. 1999. "The Breeding Migration of Marbled Salamanders (*Ambystoma opacum*) and Spotted Salamanders (*A. maculatum*) at the Jug Bay Wetlands Sanctuary on Maryland's Coastal Plain" A Technical Report of the Jug Bay Wetlands Sanctuary. Lothian, Maryland

[4] Weather Underground, www.wunderground.com

[5] Citizen Observer Weather Program, www.wxqa.com

[6] Weather Bug, www.weathebug.com

[7]Semlitsch, Raymond D., Scott, David E., Pechmann, Joseph H. K. and Gibbons, J. Whitfield, "Phenotypic variation in the arrival time of breeding salamanders: individual repeatability and environmental influences" *Journal of Animal Ecology* (1993) 62, 334-340

[8] US Naval Observatory "History of Daylight Time in the US http://aa.usno.navy.mil/faq/docs/daylight_time.html

[9] Blanchard, Frank N "The stimulus to the breeding migration of the spotted salamander *Ambystoma maculatum* (Shaw)" *The American Naturalist* (1930) 64:154-167

[10] Sexton, O. J. , Phillips, C and Bramble, J.E. "The effects of temperature and precipitation on the breeding migration of the spotted salamander (*Ambystoma maculatum*)" *Copeia*, Vol. 1990, No. 3 (Sep. 19, 1990), pp. 781-787