Ecosystem Analysis Using Probabilistic Relational Modeling

Bruce D'Ambrosio, Eric Altendorf, and Jane Jorgensen
CleverSet Inc..
{dambrosi, eric, jorgenj}@cleverset.com

Abstract

In this paper, we present the results of initial explorations into the application of relational model discovery methods to building comprehensive ecosystem models from data. Working with collaborators at the USGS Biological Resources Discipline and at the Environmental Protection Agency, we are engaged in two projects that apply relational probabilistic model discovery to building "community-level" models of ecosystems. A community-level ecosystem model is an integrated model of the ecosystem as a whole. The goal of our modeling effort is to aid domain scientists in gaining insight into data. Our preliminary work leads us to believe the method has tremendous promise. At the same time, we have encountered some limitations in existing methods. We briefly describe two projects and make some observations, particularly with respect to the development of synthetic, or derived, variables. We describe specific extensions we made to solve problems we encountered, and suggest elements of an extended grammar for such variables.

1. Introduction

Ecosystems are composed of interacting populations of organisms and their environments. They are notoriously difficult to study because of their size and complexity. In addition, many are unique. Controlled experimentation in these ecosystems is undesirable because of the potentially irreversible damage it may cause. However, observational data are often abundant. The challenge in studying ecosystems is to synthesize these data into coherent, comprehensive, biologically meaningful models.

While data collection traditions and techniques are mature, data analysis methodologies are less well developed. Generally, individual, domain-specific teams (e.g., a team of physicists or a team of biologists) apply traditional statistical methods to investigate pair-wise correlations among variables in their separate datasets, but have no methods for investigating the complex, noisy, cross-disciplinary interactions that are crucial to understanding the ecosystem as a whole. As a result, the standard ecosystem-level computational scientific method is a form of "generate and test": the manual construction of mechanistic models and model selection by comparing simulation results to data or expert knowledge. Probabilistic models of ecosystems are slowly becoming more common, however these have been constructed using knowledge-engineering (Kuikka *et al.,* 1999, Marcot *et al.,* 2001).

Most of the data collected in studies of ecological systems is stored in relational databases. An emerging family of methods for relational learning [Muggleton and De Raedt, 1994], [Van Laer and De Raedt, 2001], [Quinlan, 1996], [Getoor *et al*., 1999] provide the opportunity to learn comprehensive models directly from these relational data sources.

In this paper, we present the results of initial explorations into the application of model discovery methods to build comprehensive ecosystem models from data. Working with collaborators in the USGS Biological Resources Discipline and the Environmental Protection Agency, we are engaged in two projects that apply probabilistic relational model discovery to build "community-level" models of ecosystems. (A community-level ecosystem model is an integrated model of the ecosystem as a whole.) The goal of our modeling effort is to aid domain scientists in gaining insight into data and to construct complex prior hypotheses about the ecosystems studied. Our preliminary work leads us to believe the method has tremendous promise. At the same time, we have encountered some limitations in existing methods. We briefly describe two projects and make some observations, particularly with respect to the development of "synthetic", or derived, variables.

Probabilistic relational model discovery methods exploit a relational data model to derive parameters that account for variation in the explicit variables in a data model. In a Hollywood database, for example, an *actor's* income may be related to the *number* of *movies* in which s/he played a *role*. [Getoor *et al*., 1999] introduce the concepts of a *path* (a chain of references – e.g. "actor.role" above), and a terminal *aggregator* (e.g., "number" or count above) as defining a space of synthetic variables. We have found this framework useful, but limited in its ability to account for all known interactions in our data. We will describe examples motivating the introduction of two additional features, *selectors* and *variables*, into a synthetic variable grammar.

2. Applications

CleverSet is currently engaged in two ecological modeling projects: community-level modeling of the Crater Lake ecosystem (USGS) (Jorgensen *et al.,* 2003) and community-level modeling of West Nile virus disease transmission (Orme-Zavaleta *et al.,* 2003).

Crater Lake

Data

The National Park Service is concerned about long-term changes in the clarity of Crater Lake, a national park and the clearest deep-water lake in the world. Although many domain-specific surveys have been undertaken, the analytical framework necessary to link these analyses into one overall assessment of lake health has been lacking. Our goal in this project has been to formulate multiple, complex, simultaneous hypotheses given all the data obtained from the long-term studies of the lake (Larson *et al.*, 1993). These data have been collected using varying time and spatial scales. For example, surface weather condition information is available on a daily basis, but phytoplankton densities are measured only once or twice a month (and not at all in winter), while rocket-borne instrumentation to gather weather data at altitude is only rarely available.

Method

In an initial Crater Lake analysis performed for USGS, we chose a set of temporal units to frame the analysis. These units were time periods corresponding to observed patterns of clarity of the lake and for which data were available: June-July, August, September-October. We then added a table containing these time units (this unary relation establishes the basic time scale), and relating hydrological seasons annually (this binary relation establishes the basic unit of time-lag to be considered in the analysis), and related the data tables we wished to include in the analysis to this temporal table. A complete schema for the analysis is shown in Figure 1.

Results

Figure 2 shows the essential elements of the discovered model (we omit some schema elements for clarity). One relationship we discovered is that the dominant fish species in gill net catches was probabilistically dependent upon Secchi descending depth (water clarity) in the current year, mean fish weight in the current year, descending Secchi depth the previous year and dominant fish species two years previous. This and findings concerning age class structure agreed with the anecdotal evidence that schools of Kokanee smolts swimming at the edges of the lake were preyed upon by mature Rainbow trout, where they were caught in gill nets. This phenomenon does not occur every year. A time lag of two years, discovered by the model, is consistent with experts' observations. The relation between this interaction and water quality was previously unknown. Other somewhat surprising discoveries include: (1) the centrality of water clarity (measured by the Secchi "DesDepth" parameter); and (2) the lack of a direct relationship between Zooplankton count and water clarity, at least at the spatio-temporal scale studied. These finding suggest that fish attributes may serve as a predictor of water clarity.

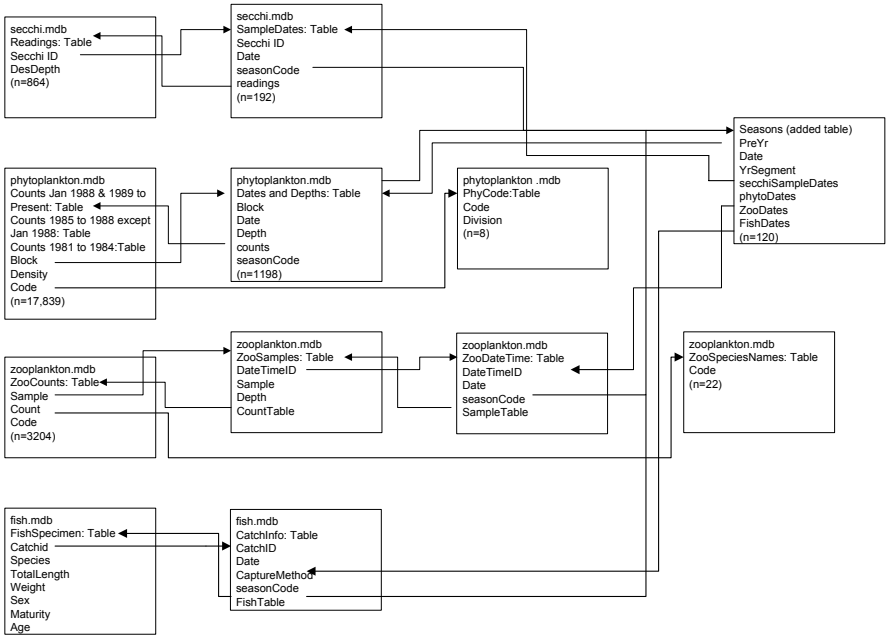

Figure 1. Crater Lake Schema

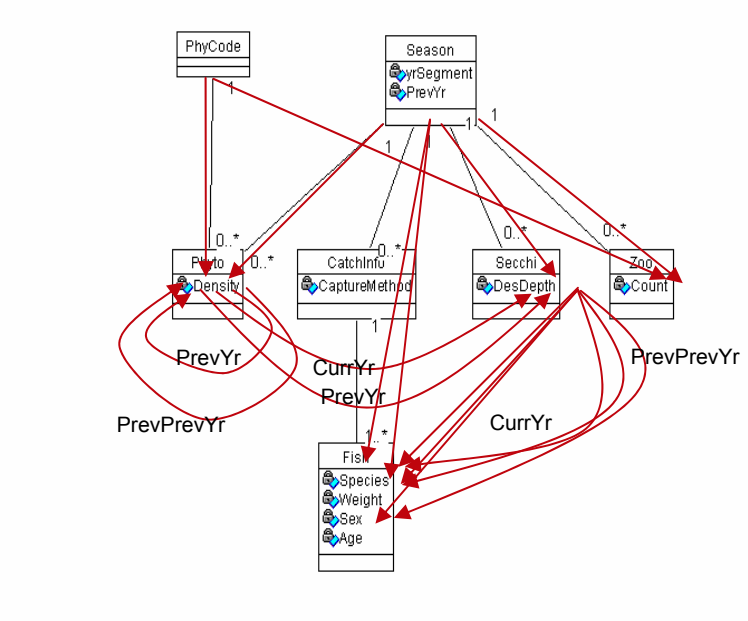

Figure 2. Crater Lake PRM

Discussion

The Crater Lake project highlighted the centrality of time in such analyses. Time creates several challenges for relational model discovery:

1. Time is rarely reified in relational schema. This presents a problem in constructing paths like "secchi.DesDepth.yrSegment.Phyto.density." Our solution in this case was to manually add a "Season" table. We have since implemented facilities for partially automating this process, by recognizing and re-ifying data/time information in schema's.
2. Once time was reified, two further decisions were necessary: we established an aggregation unit for time and we separately established a lag duration. Expert knowledge was used to establish both, based on domain knowledge and understanding of the goals of the modeling. In future we hope to explore extensions of existing statistical time series analysis methods to aid in this process.

A second problem that arose in this analysis was the frequent desire to form synthetic variables outside the scope of the current path language. For example, there were times when prior knowledge suggested that the density of a particular phytoplankton species might be a relevant parameter. Our current synthetic variable grammar does not allow for selection of a subset of the items retrieved by a path.

Finally, the goal of this project was to gain scientific insight into data that had been collected over 25 or more years (Secchi depth readings go back to the 1880s!). We found that learning models over not just the variables in the provided tables, but over their parents as well, provided additional insight. An example fragment from such an extended model, for the FishSpecimen table and its immediate parents, is shown in Figure 3. This extended model shows interactions not obvious in Figure 2, such as the multiple pathways through which Mean Secchi depth (two years previous) interacts with current Mean fish age.

West Nile Virus

Data

While the Crater Lake project involves building a relational model over multiple databases of similar type, our work with the EPA on modeling the spread of West Nile Virus involves combining multiple databases of differing types. One class of database contains incident reports (e.g., reports of dead birds testing positive for WNV, report of pools of water in which breeding mosquito populations test positive for WNV, human case reports, etc.). Each database contains reports of one type of event, located in place and time. A second class of database contains records of static features, such as the presence of a tire disposal facility (potential mosquito breeding site) or landscape type at a location. The challenge was to integrate these multiple databases into an overall model of West Nile Virus spread.

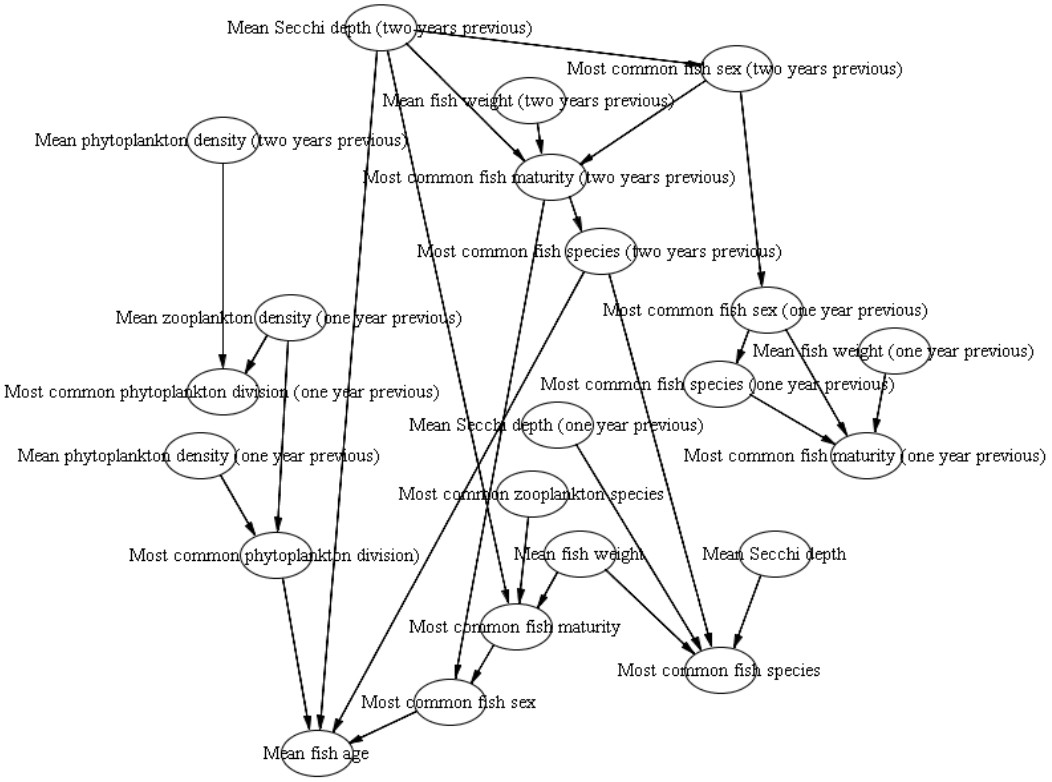

Figure 3. FishSpecimen Unrolled Model

Method

The first step in our integration of these data sources was the construction of an integrated data schema across these data sources through the addition of intensional relations linking the information in space and time. Knowing that each database recorded location in columns labeled *latitude* and *longitude*, and time as *day/month/year*, enabled us to construct a common spatio-temporal frame of reference. The simple recognition of point location in space and time, however, is not enough to integrate these data sources. Rarely do two events occur at precisely the same place or time. Rather, we imposed a *scale* across both the spatial and temporal dimensions. The parameters of this scale (five miles for space, and one month for time) were drawn from scientific knowledge about the life cycle of the vector of interest, the mosquito, and the typical flight distance for the competent bird host. Again, this was done by hand in our preliminary studies to date.

Results

Figure 4 shows a preliminary model of the spread of West Nile Virus in Maryland in 2001. Shown is a model over the synthetic variables constructed starting from the table of positive bird records.

The results support previous hypotheses that tire disposal site license density is correlated with incidence of West Nile Virus in birds. Tire disposal facilities may affect disease spread directly, by serving as breeding areas for mosquitoes, or may be a proxy for population density, which may in turn affect sampling and/or disease prevalence (e.g., though human movement through the region). The results also suggest that disease prevalence in mosquito pools may be a predictor of disease appearance in birds. The number of human and horse cases in 2001 was too small to support any significant findings related to these cases. However, even with these sparse data, the model produced is consistent with current knowledge regarding the manner in which the disease is transmitted and forms a framework in which future findings may be evaluated. The fact that horse cases do not contribute significant information to the model provides preliminary evidence that monitoring this incompetent host may be unnecessary in tracking the spread of this disease.

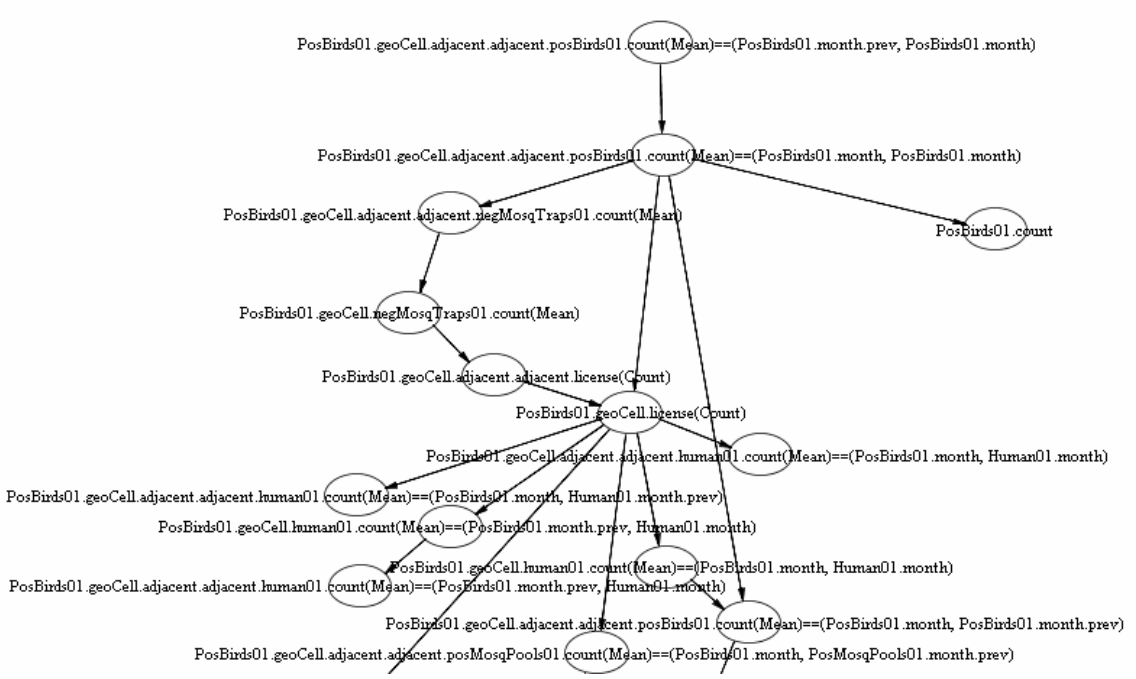

Figure 4. West Nile Virus Model Fragment

Since the mechanistic model of disease spread is not completely known, the temporal and spatial models included in the model may not be the only, or even the most useful scales at which to view interactions. Finer spatial resolutions, for example, might provide evidence about the species of birds and mosquitoes involved in transmission. Landscape level data, for example, landcover type, might also improve the descriptive and predictive capabilities of the model. As mentioned in our discussion of the Crater Lake study, our current manual methods do not permit easy exploration of possible scales.

Discussion

Our work on West Nile Virus propagation reinforces the need for selectors in synthetic variables. Unlike Crater Lake, however, where the selectors where over the values of primitive attributes, in the analysis of West Nile Virus, we needed to form equality selectors over entities (e.g., positive mosquitoes in adjacent geocells *in the same month*). We extended our synthetic variable grammar to include a single *selector* phrase. A selector is a Boolean operator mapped over the elements of the base path defining a synthetic variable. Elements for which the selector returns *true* and included in the result, and elements for which it returns *false* are omitted. The selector consists of a Boolean operator and two paths. The first path is applied to the table entry at the head of the base path for the synthetic variable, and the second path is applied to each table entry retrieved by the base path. For example, consider:

PosBirds.GeoCell.PosMosq ==(PosBird.month, PosMosq.month).Count()

The base path ("PosBirds.GeoCell.PosMosq") yields a set of positive mosquito entries in the same spatial region as a bird entry. The selector ("==(PosBird.month, PosMosq.month)") then filters out all entries not in the same month as the positive bird record. Finally, the "Count()" aggregator returns a scalar, the cardinality of the resulting set[1].

3. Conclusions and Future Work

Relational probabilistic modeling provides a natural framework for investigating ecological data. The large amount of observational, noisy data, often collected by multiple investigators over varying time-scales, provides a rich field for probabilistic model discovery, and relational approaches raise the level of modeling to one with which domain scientists can readily interact.

Existing synthetic variable construction methods naturally generate many variables either previously

[1] In more recent work, supported by NSF SBIR DMI-0231961, we have developed a more comprehensive synthetic variable language grammar and automated generation capability, patent-pending.

known to scientists or immediately recognized by them as scientifically relevant. At the same time, attempts to apply relational probabilistic model discovery techniques to ecological data have revealed limitations in our current synthetic variable construction methods. We are currently exploring work in data base *path expressions*, for example that of Van den Bussche [Van den Bussche *et al.*, 93] and Frohn [Frohn *et al.*, 94], as generalizations capable of expressing a more comprehensive set of synthetic variables. Key concepts include the *selector* and the introduction of *variables* (to allow subsequent reference to earlier elements in a path). We are also exploring mixed-initiative search procedures over these much larger path grammars.

Acknowledgements

The authors wish to acknowledge financial support from the USGS for some of the work reported here. We also thank Gary L. Larson, Research Manager, USGS Forest and Rangeland Ecosystem Science Center, Corvallis, OR and Jennifer Orme-Zavaleta, Associate Director for Science, USEPA/
NHEERL/WED, Corvallis, OR for their involvement and support.

References

[Frohn *et al*., 1994] Frohn J., Lausen G., Uphoff H., "Access to objects by path expressions and rules", Proceedings of 20th International Conference on Very Large Databases, 1994.
http://citeseer.nj.nec.com/frohn94acces.html

[Getoor *et al*., 1999] Getoor, L., N. Friedman, D. Koller, and A. Pfeffer. 1999. Learning probabilistic relational models. In *Proceedings of Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-1999),* Stockholm, Sweden. http://citeseer.nj.nec.com/context/889848/71217

[Jorgensen *et al.,* 2003] Jorgensen, J., B. D'Ambrosio, and P. A. Rossignol. 2003. Data-Driven Construction of Community Models of Crater Lake. *NSF Biocomplexity Workshop -The vertical organization of energy, carbon, and nutrient cycles in an ultraoligotrophic ecosystem: A workshop on Crater Lake, Oregon*. February 16 – 18, 2003.

[Kuikka *et al*., 1999] Kuikka, S , M. Hilden, H. Gislason, S. Hanson, H. Sparholt, and O. Varis. 1999. Modeling environmentally driven uncertainties in Baltic cod (*Gadus morhua*): Management by Bayesian influence diagrams. *Canadian Journal of Fisheries and Aquatic Sciences*. 54:629-641.

[Larson *et al*., 1993] Larson, G. L., C. D. McIntire and R.W. Jacobs, editors. 1993. Crater Lake Limnological Studies Final Report. National Park Service, Seattle, WA. 722 pp.

[Marcot *et al*., 2001] Marcot, B. G., R. S. Holthausen, M. G. Raphael, M. M. Rowland, and M. J. Wisdom. 2001. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management* 153:29-42.

[Muggleton and De Raedt, 1994] Muggleton, S. and L. De Raedt. 1994. Inductive logic programming: Theory and methods. Journal of Logic Programming, 20: 629-679.

[Orme-Zavaleta *et al.,* 2003] Orme-Zavaleta, J., J. Jorgensen, B. D'Ambrosio, H-K. Luh and P. A. Rossignol. 2003. Data-driven discovery of temporal and geospatial patterns of disease transmission: West Nile Virus in Mayland. *National Conference on USGS Health-Related Research – Natural Science and Public Health: Prescription for a better Environment*. April, 2003.

[Quinlan, 1996] Quinlan, J. R. 1996. Learning first-order definitions of functions. *Journal of Artificial Intelligence Research,* (October) 5:139-161.

[Van den Bussche and Gottfried, 1993] Jan Van den Bussche and Gottfried Vossen. An extension of path expressions to simplify navigation in object-oriented queries. In Stefano Ceri, Katsumi Tanaka, and Shalom Tsur, editors, Deductive and ObjectOriented Databases, pages 267--282, Phoenix, Arizona, 1993. Springer Verlag, Lecture Notes in CS, No. 760. http://citeseer.nj.nec.com/vandenbussche93extension.html

[Van Laer and De Raedt, 2001] Van Laer, W. and L. De Raedt. 2001. How to upgrade propositional learners to first order logic: Case study. *Lecture Notes in Computer Science*, 2049:102.