Regression by hybrid Bayesian networks: modelling landscape - socioeconomic relationships

Rosa F. Ropero\textsuperscript{1}  P. A. Aguilera\textsuperscript{1}
Antonio Fernández\textsuperscript{2}  Rafael Rumí\textsuperscript{2}

\textsuperscript{1}Informatics and Environment Research Group, Department of Biology and Geology, University of Almería
\textsuperscript{2}Department of Mathematics, University of Almería

Albacete, February 7-8, 2013
Motivation

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries

Hybrid BNs-based regression

Modelling landscape - socioeconomic relationship

Conclusions

Landscape ↔ socioeconomic
Motivation (Aguilera et al. 2011)

- Modelling environmental problems with continuous (and discrete) variables is a challenge.
- Modelling landscape - socioeconomic relationships with BNs.
- Advantages using BN-based regression.
- In Almería this study is even more interesting:
  - Varied landscape (desert/high mountains/sea, traditional crops/greenhouses, ...)
  - Changing landscape (boom of greenhouses, ...)
1 Preliminaries

2 Hybrid BNs-based regression

3 Modelling landscape - socioeconomic relationship

4 Conclusions
Preliminaries

Hybrid BNs-based regression

Modelling landscape - socioeconomic relationship

Conclusions
Hybrid Bayesian networks

Definition

A BN is called **hybrid** when continuous and discrete variables coexist simultaneously in the model.
Hybrid Bayesian networks

Definition

A BN is called **hybrid** when continuous and discrete variables coexist simultaneously in the model.

Approaches dealing with hybrid data

- Discretisation.
- Conditional Linear Gaussian (CLG) model.
- Mixtures of Truncated Exponentials (MTEs).
- Mixtures of Polynomials (MOPs).
- Mixtures of Truncated Basis Functions (MoTBFs).
Mixtures of Truncated Exponentials (MTEs)

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries

Hybrid BNs-based regression

Modelling landscape - socioeconomic relationship

Conclusions

Mixtures of Truncated Exponentials (MTEs)

\[ f(z) = \begin{cases} 
-0.0172 + 0.931e^{1.27z} & \text{if } -3 \leq z < -1 \\
0.442 - 0.0385e^{-1.64z} & \text{if } -1 \leq z < 0 \\
0.442 - 0.0385e^{1.64z} & \text{if } 0 \leq z < 1 \\
-0.0172 + 0.9314e^{-1.27z} & \text{if } 1 \leq z < 3 
\end{cases} \]

\[ P(Y = 1|z) = \begin{cases} 
0 & \text{if } z < -5 \\
-0.0217 + 0.522e^{0.635z} & \text{if } -5 \leq z < 0 \\
1.0217 - 0.522e^{-0.635z} & \text{if } 0 \leq z \leq 5 \\
1 & \text{if } z > 5 
\end{cases} \]
Contents

1 Preliminaries

2 Hybrid BNs-based regression

3 Modelling landscape - socioeconomic relationship

4 Conclusions
Regression

Assume a set of variables $Y, X_1, \ldots, X_n$, where $Y$ is the dependent variable and $X_i$ are the independent ones.

Goal

Find a model $g$ to explain $Y$ in terms of $X_1, \ldots, X_n$:

$$y = g(x_1, \ldots, x_n)$$
**BN-based regression**

\( g \) can be modelled by a BN: \( f(y \mid x_1, \ldots, x_n) \).

**Classification**

An individual with observed features \( x_1, \ldots, x_n \) will be classified as a member of class \( y^* \) obtained as:

\[
y^* = \arg \max_{y \in \Omega_Y} p(y \mid x_1, \ldots, x_n) .
\]

**Regression**

\( f(y \mid x_1, \ldots, x_n) \) is computed and a numerical prediction for \( Y \) is given using the expected value as follows:

\[
\hat{y} = \mathbb{E}[Y \mid x_1, \ldots, x_n] = \int_{\Omega_Y} y f(y \mid x_1, \ldots, x_n) dy
\]
Regression with a NB structure

Let $Y$ be the **continuous** dependent variable and $X_1, \ldots, X_n$, the discrete/continuous independent variables.

\[
f(y \mid x_1, \ldots, x_n) \propto f(y) \times f(x_1, \ldots, x_n \mid y)
\]

\[
\propto f(y) \prod_{i=1}^{n} f(x_i \mid y),
\]
Feature selection (Morales et al. 2007)

Filter-wrapper approach

1. Features are sorted using as filter measure $I(X, Y)$:

\[
I(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \log_2 \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)} \, dx \, dy
\]

2. Using that order, they are included one by one whenever it increases the accuracy of the model:

\[
rmse = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

\[
\hat{I}(X, Y) = \frac{1}{m} \sum_{i=1}^{m} \left( \log_2 f_{X|Y}(X_i \mid Y_i) - \log_2 f_X(X_i) \right)
\]
Feature selection

The variable selection is performed during learning

Sorted features by $\hat{I}(X_i, Y) : X_2, X_3, X_1, X_4$

$\text{rmse} = 0.15$

$Y \xrightarrow{\text{Add } X_3} X_2$

$\text{rmse} = 0.14$

$Y \xrightarrow{\text{Add } X_1} X_2 \xrightarrow{\text{Add } X_3} X_1$

$\text{rmse} = 0.145$

$Y \xrightarrow{\text{Add } X_4} X_2 \xrightarrow{\text{Add } X_3} X_4$

$\text{rmse} = 0.143$

$Y \xrightarrow{\text{Remove } X_1} X_2 \xrightarrow{\text{Add } X_4} X_3 \xrightarrow{\text{Remove } X_4} X_2 \xrightarrow{\text{Add } X_3} X_4$

$\text{rmse} = 0.14$

FINAL MODEL

$Y \xrightarrow{\text{Remove } X_4} X_2 \xrightarrow{\text{Add } X_3} X_2$
## Contents

1. Preliminaries
2. Hybrid BNs-based regression
3. Modelling landscape - socioeconomic relationship
4. Conclusions
Study area (southeastern Spain)

- **Regions**: Poniente almeriense, Sierra de los Filabres and Alpujarra.
- **90 municipalities, 500,000 Has.**
- Landscape characterized by an **altitude gradient**.
- **Low** part: greenhouses, high population, immigration.
- **Middle-high** part: agricultural mediterranean landscape, low population, emigration, rural tourism.
Data collection

- Data per municipality in 2007.
- Sources: SIMA (Sistema de Información Multiterritorial de Andalucía) and REDIAM (Consejería de Medio Ambiente).

Landscape tendencies (3)
- Increase of scrublands.
- Agricultural mediterranean landscape.
- Native forest.

Socioeconomic variables (16)
- Total population, Aging, Natural increase, Rate of male.
- Primary/Secundary/Tertiary sector, Unemployed.
- National/Foreign emigration, National/Foreign immigration.
- Illiterate, Primary/Secondary/Higher studies.
Scenarios of change *(Schmitz et al. 2005, Aranzabal et al. 2008)*

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Variables</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive socioeconomic change</strong></td>
<td>Foreign immigration</td>
<td>Maximum value</td>
</tr>
<tr>
<td></td>
<td>National emigration</td>
<td>+50%</td>
</tr>
<tr>
<td></td>
<td>Tertiary sector</td>
<td>+60%</td>
</tr>
<tr>
<td></td>
<td>Primary sector</td>
<td>+80%</td>
</tr>
<tr>
<td></td>
<td>Higher studies</td>
<td>+15%</td>
</tr>
<tr>
<td></td>
<td>Secondary studies</td>
<td>+30%</td>
</tr>
<tr>
<td></td>
<td>Natural increase</td>
<td>+70%</td>
</tr>
<tr>
<td></td>
<td>Aging</td>
<td>Minimum value</td>
</tr>
</tbody>
</table>
### Scenarios of change (Schmitz et al. 2005, Aranzabal et al. 2008)

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Variables</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive socioeconomic change</strong></td>
<td>Foreign immigration</td>
<td>Maximum value</td>
</tr>
<tr>
<td></td>
<td>National emigration</td>
<td>+50 %</td>
</tr>
<tr>
<td></td>
<td>Tertiary sector</td>
<td>+60 %</td>
</tr>
<tr>
<td></td>
<td>Primary sector</td>
<td>+80 %</td>
</tr>
<tr>
<td></td>
<td>Higher studies</td>
<td>+15 %</td>
</tr>
<tr>
<td></td>
<td>Secondary studies</td>
<td>+30 %</td>
</tr>
<tr>
<td></td>
<td>Natural increase</td>
<td>+70 %</td>
</tr>
<tr>
<td></td>
<td>Aging</td>
<td>Minimum value</td>
</tr>
<tr>
<td><strong>Negative socioeconomic change</strong></td>
<td>National emigration</td>
<td>Maximum value</td>
</tr>
<tr>
<td></td>
<td>Higher studies</td>
<td>−70 %</td>
</tr>
<tr>
<td></td>
<td>Natural increase</td>
<td>Minimum value</td>
</tr>
<tr>
<td></td>
<td>Primary sector</td>
<td>−20 %</td>
</tr>
<tr>
<td></td>
<td>Tertiary sector</td>
<td>−80 %</td>
</tr>
<tr>
<td></td>
<td>Total population</td>
<td>−50 %</td>
</tr>
<tr>
<td></td>
<td>Aging</td>
<td>+80 %</td>
</tr>
<tr>
<td></td>
<td>Secondary studies</td>
<td>−40 %</td>
</tr>
<tr>
<td></td>
<td>Unemployment</td>
<td>Maximum value</td>
</tr>
</tbody>
</table>
Agricultural mediterranean landscape (AML)

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries
Hybrid BNs-based regression
Modelling landscape - socioeconomic relationship
Conclusions
Agricultural mediterranean landscape (AML)

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries
Hybrid BNs-based regression
Modelling landscape - socioeconomic relationship
Conclusions

Agricultural Mediterranean Landscape

<table>
<thead>
<tr>
<th>Probability</th>
<th>A priori</th>
<th>Positive scenario</th>
<th>Negative scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Expected value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A priori</td>
<td>21.50</td>
</tr>
<tr>
<td>Scenario +</td>
<td>15.62</td>
</tr>
<tr>
<td>Scenario -</td>
<td>29.65</td>
</tr>
</tbody>
</table>
Native forest
Native forest

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries

Hybrid BNs-based regression

Modelling landscape - socioeconomic relationship

Conclusions

Native forest

Natural increase

National immigration

National emigration

Primary studies

Tertiary sector

Total population

Expected value

A priori

Scenario +

Scenario -

Probability

Native forest

0.00 0.05 0.10 0.15

0 10 20 30 40 50

A priori

Positive scenario

Negative scenario

A priori 7.66

Scenario + 3.81

Scenario - 6.32
Scrubland

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries

Hybrid BNs-based regression

Modelling landscape - socioeconomic relationship

Conclusions
Scrubland

Regression by hybrid BNs: modelling landscape - socioeconomic relationship

R. F. Ropero, P. A. Aguilera, A. Fernández, R. Rumí

Preliminaries
Hybrid BNs-based regression
Modelling landscape - socioeconomic relationship
Conclusions

Scrubland

Aging
Unemployed
Tertiary sector
Foreign emigration
National emigration
Secondary studies
Primary studies

A priori
Positive scenario
Negative scenario

Expected value

A priori 39.62
Scenario + 26.96
Scenario - 49.98
Validation

2 approaches

- Multiple Linear Regression (MLR) included in R.
- Approach with the proposed model (BN).

\[
\text{rmse} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

- 10 fold-cross-validation is used to reduce variability.

<table>
<thead>
<tr>
<th></th>
<th>Native forest</th>
<th>AML</th>
<th>Scrubland</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>5.20</td>
<td>18.01</td>
<td>19.09</td>
</tr>
<tr>
<td>MLR</td>
<td>8.81</td>
<td>19.92</td>
<td>29.47</td>
</tr>
</tbody>
</table>
Contents

1 Preliminaries

2 Hybrid BNs-based regression

3 Modelling landscape - socioeconomic relationship

4 Conclusions
Concluding remarks

- The work presents MTE-based hybrid BNs as a tool for solving regression problems in environmental sciences.

<table>
<thead>
<tr>
<th>Advantages in environmental sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous and discrete variables are allowed without restriction on the structure.</td>
</tr>
<tr>
<td>Any underlying data distribution is permitted.</td>
</tr>
<tr>
<td>Probabilistic information (not only for the dependent variable).</td>
</tr>
<tr>
<td>Partial evidences.</td>
</tr>
<tr>
<td>A wider range of problems can be addressed.</td>
</tr>
</tbody>
</table>
Concluding remarks

- 3 landscape tendencies and several socioeconomic aspects have been modeled.

**2 change scenarios analysed**

- **Scenario +**: Traditional landscapes are replaced by touristic activities.
- **Scenario -**: It implies an increase of scrublands, and the maintenance of traditional croplands in rural areas.

**Future works**

Data dependence in environmental problems.
Thanks for your attention