A Minimum Crossed Entropy Model to Generate Information at Disaggregated Level

RUI FRAGOSO
Departamento de Gestão de Empresas
Universidade de Évora
Lg. dos Colegiãis, 1 – 7000 ÉVORA
PORTUGAL
rfragoso@uevora.pt

MARIA DE BELÉM MARTINS
Faculdade de Engenharia dos Recursos Naturais
Universidade do Algarve
Campus de Gambelas, 8005-139 FARO
PORTUGAL
mbmartins@ualg.pt

MARIA RAQUEL VENTURA LUCAS
Departamento de Gestão de Empresas
Universidade de Évora
Lg. dos Colegiãis, 1 – 7000 ÉVORA
PORTUGAL
mrlucas@uevora.pt

Abstract: - Montado ecosystem in the south of Portugal, integrated in Mediterranean ecosystems, has been extensively used by man for agro-forestry-extensive grazing systems. Sustainable development of this Montado ecosystem demands a long-term perspective about the consequences of today’s activities, but a definition of the homogeneous units among this heterogeneous ecosystem is absolutely needed. This paper demonstrates the adherence of the minimum crossed entropy model as a process conducting to dynamic spatial information generation and disaggregation. With this model, it will be possible to use the disposable information to generate data disaggregated to homogeneous units that can be used on the estimation of these territories’ soil occupation. The results quality will be evaluated using the weighted prescription absolute deviation (DAPP) and measuring the quantity of information that is changed due to aggregation process by the use of a “Disaggregation Information Gain” (DIG), based on the cross entropy between the observed values of soil occupation at aggregated and disaggregated level and on the crossed entropy between the soil occupation estimated by the disaggregation model and those observed at HSU disaggregated level.

Key-Words: Montado ecosystem, minimum crossed entropy model, homogeneous units, Alentejo.

1 Introduction
Montado ecosystem in the Alentejo Region, south of Portugal, integrated in Mediterranean ecosystems, has been extensively used by man for agro-forestry-extensive grazing systems. Due to its extensive use, this ecosystem is not severely impacted, even with strong ecological restrictions like long summer drought. Oak forests and the savanna-like landscapes, where Quercus ilex spp. rotundifolia and Quercus suber are a dominant part of the agro forestry system, and produce fodder for livestock, as well as cork and firewood.

Sustainable development of this Montado ecosystem demands a long-term perspective about the consequences of today’s activities. It goes beyond economic aspects to include environmental and social concerns and the recognition that total co-operation is required to achieve quantitative and qualitative information to support analytical and methodological approaches to enable sustainable conditions. The Montado ecosystem area (MEA) is far from being homogeneous; there are enormous and varied heterogeneities, namely in what concerns agro-ecological aspects, that determine the specialization
profile of economic activities, particularly the agro-forestry activities. The management of such a heterogeneous ecosystem is very difficult. For this reason, the determination of homogeneous units that allow the heterogeneous macro magnitudes of the MEA characterization is of particular importance. This paper proposal is to demonstrate the adherence of the minimum crossed entropy model as a process conducting to dynamic spatial information generation and disaggregation. With this model, it will be possible to use the disposable information to generate data disaggregated to homogeneous sub-units (HSU) that can be used on the estimation of these territories’ soil occupation.

2 Homogeneous territorial units of the Montado Ecosystem Area

The identification of homogeneous territorial units of MEA was based on the ecosystems classification using non-hierarchical multivariate analysis and non-linear methods. The criteria used implied the study of representative dominant patterns, based on the following indicators for all the Alentejo parishes:

1 – Climatic averages, maximum temperature in August, minimum temperature in January, inland or not, Winter precipitation, Summer precipitation and altitude.
2 – Soil allocation in what concerns dominant agro-forestry use (agriculture, forestry or not cultivated).
3 – Average farm dimension.
4 – Type of Montado: Quercus ilex spp. rotundifolia with low density, Quercus ilex spp. rotundifolia with high density, Quercus ilex spp. rotundifolia with very high density, Quercus suber with low density, Quercus suber with high density, Quercus suber with very high density, mixed with low density, mixed with high density, mixed with very high density and Pinus pinea with low density.
5 – Type of cattle: head density and proportion of farms with sheep, swine, beef cattle and goats.

For each of these five groups of variables a first analysis was made, reducing the information to the two variables that comprehend more then 80% of the group variance. Then parishes were classified in six classes with a K-means analysis and this classification was analysed through an individual classification tree to determine which variables characterized each class. The results of K-means analysis identified six classes, corresponding to six dominant patterns, that allowed the establishment of six homogeneous agro-forestry macro-units in MEA. Each of these homogeneous macro-units is linked with the diversity and regional specificities of available resources as well as how these resources are used and valued by resident populations.

Climatic factors, soil type, slopes, forest characteristics and littoral proximity define structural characteristics of natural resources. In general, these factors determine the resources’ use and evaluation and play a major role on the populations’ way of living.

The natural resources’ potential depends mainly on the way their limiting characteristics interact with the factors labour, capital and entrepreneur to attain a certain objective. The results of this interaction result in an economic dimension and different types of soil allocation and cattle breeding technologies.

The six homogeneous macro-units identified are shown in the annex, table 1. These homogeneous macro-units were then crossed with environmental protection areas of NATURA 2000 and National Protected Areas and 31 homogeneous polygons were defined, as a function of their interest for natural values conservation, which became the 31 homogeneous sub-units (HSU) that can be seen in the annex, table 2.

3 The dynamic process of allocating soil inside the HSU

Having defined the pertinent HSU, the problem is now to estimate the soil occupation and breeding cattle activities for each HSU, based on the available data for MEA and on the disaggregated available information for years 1989 and 1999. This is particularly important, since the identification of regional models that integrate in the MEA sustainable agro-forestry activities supposes the evaluation of each HSU agro-forestry potential. The simple definition of the HSU partially responds to this evaluation, but it only considers the potential use of resources and not the effective use of resources, that represents the organization, in the MEA, of technical, economic and institutional factors.

It is possible to specify soil occupation and breeding cattle at MEA aggregated level for years 1989, 1999 to 2004 from 1989 and 1999 Agricultural Census [1], Structural Agricultural Enterprises Surveys of 2003 and 2005 [2, 3] and Economic Agricultural Accounts from 1999 to 2005 [4, 5]. The disaggregated results for each HSU can also be obtained to 1989 and 1999 and its estimation for the series’ remaining years until 2004 can be done adopting a method based on the problem of spatial disaggregation to estimate incomplete information.

The methodology used, like in Howitt et Reynaud [6], differs from the one of Miller et Plantinga [7] because
the process of soil occupation choice is endogenous. So, the question is to combine in a dynamic process disaggregated incomplete information and aggregated complete information to fulfill missing information.

Considering there are $I$ HSU and that soil occupation in each year is given by $S_k(t)$, where $k=1,...,K$ corresponds to observed agro-forestry activities and $t=1,...,T$ corresponds to the year in which they occur, then the probability of producing $k$ activity in year $t$ is:

$$\gamma_k(t) = \frac{S_k(t)}{\sum_{k=1}^{K} S_k(t)} \quad \forall k, t = 1,...,r$$

(1)

The disaggregated information at HSU level in what concerns soil occupation by each agro-forestry activity $s_k^j(t)$ is available only for the first $r$ periods ($r < T$) and it is given by the probability of producing the activity $k$ in year $t$ and HSU $i$:

$$\gamma_k(t) = \frac{s_k^j(t)}{\sum_{k=1}^{K} s_k^j(t)} \quad \forall k, t = 1,...,r$$

(2)

The estimation of $s_k^j$ must obey the following restriction that guarantees that soil allocation for each activity at MEA level is equal to the sum of this activity area in each HSU:

$$S_k(t) = \sum_{i=1}^{I} s_k^j(t) \quad \forall k, t = r+1,...,T$$

(3)

According to Kijima [8] a sequence of $r$ observations of soil allocation can be characterized by a $1^{st}$ order Markov process. In this case, any activity choice process, among $K$ possible activities for the HSU $i$ is defined as a $1^{st}$ order Markov process in the space $\{1,...,K\}$. This means that $K'$ states of decision are considered, corresponding to $K'$ possible strategies, indexed by $j$ in $\{1,...,J\}$ with $J=K'$. The probability associated to each state $j$ in the HSU $i$ and year $t$ is given by $q_k^j(t)$. $q_k^j(t)$ corresponds to the product of probabilities $\gamma_k(t)$ to the sequence of states $j$. If there was data, a $2^{nd}$ order Markov process could be assumed. As the information available at HSU level only corresponds to the first $r$ observations, the following proceeding has to be taken:

1. Estimate the aggregate probabilities transition matrix for dynamic allocation of agro-forestry activities at MEA level based on the theory of Generalized Maximum Entropy (GME) [9]

2. Disaggregation of soil occupation at HSU level through the Minimum Crossed Entropy (MCE) and using the aggregate probabilities transition matrix. The GME problem can be formulated as:

$$\max H(T,e) = \sum_{j'=1}^{J} \sum_{j=1}^{J} \sum_{t=1}^{T} T_{jj'} \log(T_{jj'}) + \sum_{k=1}^{K} \sum_{n=1}^{N} e_{kn}(t) \log(e_{kn}(t))$$

subject to:

$$\sum_{j=1}^{J} T_{jj'} = 1 \quad \forall j', j \in \{0, 1\}$$

$$\sum_{n=1}^{N} e_{kn}(t) = 1 \quad \forall j', t \in \{0, 1\}$$

subject to:

$$q_k^j(t) = \sum_{j=1}^{J} \sum_{t=1}^{T} T_{jj'} \gamma_k(j') + \sum_{n=1}^{N} e_{kn}(t) \gamma_k(t) \quad \forall k, t = r+1,...,T$$

(4)

$$\sum_{j=1}^{J} T_{jj'} = 1 \quad \forall j', j \in \{0, 1\}$$

$$\sum_{n=1}^{N} e_{kn}(t) = 1 \quad \forall j', t \in \{0, 1\}$$

subject to:

$$s_k^j(t) = \sum_{i=1}^{I} S_k(t) \gamma_k(t) \quad \forall k, t = 1,...,r$$

(5)

$$\sum_{j=1}^{J} T_{jj'} = 1 \quad \forall j', j \in \{0, 1\}$$

(6)

$$\sum_{n=1}^{N} e_{kn}(t) = 1 \quad \forall j', t \in \{0, 1\}$$

(7)

The objective of this optimization problem is to maximize the entropy of probabilities distribution $\{T_{jj'^{t1}},...,T_{jj'^{TM}}\} \forall j, j'$ and $\{e_{j'^{t1}},...,e_{j'^{tN}}\} \forall j'^{t}$, considering the restrictions imposed by equations 5 to 8.

After this, it is still need to estimate for each year probabilities transition matrix at HSU disaggregated level, solving a problem of MCE, and using the estimated transition probabilities to finally obtain soil occupation for each HSU.

The MCE problem at HSU level can be formulated as:

$$\min H(T,e) = \sum_{j'=1}^{J} \sum_{j=1}^{J} \sum_{t=1}^{T} T_{jj'} \log(T_{jj'}) + \sum_{k=1}^{K} \sum_{n=1}^{N} e_{kn}(t) \log(e_{kn}(t))$$

subject to:

$$s_k^j(t) = \sum_{i=1}^{I} S_k(t) \gamma_k(t)$$

(9)

$$\sum_{j=1}^{J} T_{jj'} = 1 \quad \forall j', j \in \{0, 1\}$$

(10)

$$\sum_{n=1}^{N} e_{kn}(t) = 1 \quad \forall j', t \in \{0, 1\}$$

(11)

$$\sum_{j=1}^{J} T_{jj'} = 1 \quad \forall j', j \in \{0, 1\}$$

(12)

where $\{\zeta_1,...,\zeta_N\}$, $N \geq 2$ is the support vector associated to probabilities $\{e_1,...,e_N\}$ such that $e_k = \sum_{n=1}^{N} e_{kn} \forall k$.

This model minimizes the cross entropy of the transition probabilities distribution and the entropy of the errors distribution probability.

On the disaggregation process from MEA level to HSU level, is necessary to guarantee that the information is compatible at different disaggregation levels. This is done using equation 10. In the model, the first member of the equation is computed each year, as a function of the modifications that occur in the agro-forestry activities allocation.

4 Results

The results should be compared with the reality, which can only be done to the year 1999, as for the other years
there are no disaggregated results. It must be stated that the estimations obtained with this model were close to the reality in 1999. The main differences are in what concerns permanent pastures and fallow. Nevertheless, these differences do not put in doubt the validity of method. In reality, many of the permanent pastures are integrated in very long rotations and subject to very low number of heads per ha and the fallows are frequently used as spontaneous pastures. So, the errors can be accommodated in current practices that can not be reflected in the disposable statistical information.

To evaluate the results quality we also used the weighted prescription absolute deviation (DAPP), calculated as follows, for each activity and HSU:

\[
DAPP_i = \frac{1}{y_k} \left| y_k - \hat{y}_k \right|
\]

and at MEA aggregated level:

\[
DAPP = \frac{\sum_k i^2 DAPP_i}{\sum_k i^2}
\]

The DAPP values by activity were relatively low. Values above 15\%, which is the value considered by Hazell et al. [10] as the reasonable calibration threshold, only occur for permanent pastures and, even thought, only for 7 of the 31 HSU. Analysing the total DAPP for each HSU, both considering all the activities or all but permanent pastures and fallows, which are the activities that can be confused, as it was said before, values of 27.3\% and 10.5\% were obtained, which is perfectly accepted as an estimation error and confirm the good method’s adherence to disaggregate information.

The second interesting question is what is the information gain we have with this process, for which we need an indicator that measures the quantity of information that is changed due to aggregation process. Howitt et Reynaud [6] construct a “Disagegation Information Gain” (DIG), based on the cross entropy between the observed values of soil occupation at aggregated and disaggregated level and on the crossed entropy between the soil occupation estimated by the disaggregation model and those observed at HSU disaggregated level.

DIG measures the proportion of heterogeneity at HSU level that is covered by the disaggregation process used. In a perfect disaggregation, DIG is equal to 1. In this case, the disaggregation process recovers 100\% of information heterogeneity. If there is no heterogeneity in the information the DIG is 0. Our model obtained a DIG of 0.43, which means the disaggregation recovers 43\% of the information heterogeneity at the 31 HSU level, considering the 14 agro-forestry activities for soil occupation. These results are very satisfactory, especially when compared to the DIG values between 56 to 69\% obtained by Howitt et Reynaud [6] in an information disaggregation process for 6 districts of California State (USA) and 8 agriculture activities.

5 Conclusions

The model proposed is able to disaggregate the values that exist at MEA level to HSU level, allowing the representation of soil allocation each year, for each homogeneous sub-unit. Economic sustainability is an important part of the sustainability issue and the economic analysis is surely the framework of the structural policy analysis that impacts the territory and of the future scenarios’ study, technical and economic management models and evaluation of determinant factors.

It is of fundamental interest to have a tool that allows the information generation for homogeneous sub-units, giving the basis to study socio-economic parameters and the impact of agricultural policy scenarios on the natural resources exploitation.

References:
Table 1 Alentejo representative Montado Agro-Forestry Production Systems (MAPS)

<table>
<thead>
<tr>
<th>MAPS</th>
<th>Soil occupation</th>
<th>Animal activity</th>
<th>Forest characteristics</th>
<th>Climatic factors</th>
<th>Agricultural economy aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Pastures under trees</td>
<td>Cattle and swine</td>
<td>Weak forest. Predominance of Quercus ilex spp. rotundifolia</td>
<td>Inner zones, with low level of precipitation</td>
<td>Big farms: 1366 ha UAS and 6 AWU</td>
</tr>
<tr>
<td>B</td>
<td>Olive oil and vineyards systems and gras and grazings activities, under trees or not</td>
<td>Cattle and sheep</td>
<td>Quercus ilex spp. rotundifolia with high density</td>
<td>Littoral zones with good level of precipitation</td>
<td>Medium to big farms: 798 ha UAS and 4,21 AWU</td>
</tr>
<tr>
<td>C</td>
<td>Grass and grazings activities, under trees</td>
<td>Cattle and swine</td>
<td>Quercus ilex spp. rotundifolia and Quercus suber with high density</td>
<td>High inner zones with good level of precipitation</td>
<td>Medium to big farms: 107 ha UAS and 1,15 AWU</td>
</tr>
<tr>
<td>D</td>
<td>Cereals and pastures, under trees or not</td>
<td>Cattle</td>
<td>Quercus suber with high density</td>
<td>Inner zones near littoral with good level of precipitation in Winter</td>
<td>Small to medium farms: 448 ha UAS and 3,05 AWU</td>
</tr>
<tr>
<td>E</td>
<td>Cereals and pastures, under trees or not</td>
<td>Cattle</td>
<td>Quercus ilex spp. rotundifolia and Quercus suber with high density</td>
<td>Inner zones with low level of precipitation</td>
<td>Small to medium farms: 135322 6,7 Elvas, Fronteira e Monforte</td>
</tr>
</tbody>
</table>

Source: k-means analysis, done by the Évora University’s Unit of Macroecology and Conservation

Table 2 – Distribution of homogeneous sub-units (HSU) by MEA

<table>
<thead>
<tr>
<th>Macro Unit</th>
<th>ha</th>
<th>%</th>
<th>Homogeneous sub-unit</th>
<th>ha</th>
<th>%</th>
<th>Predominant County</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>374686</td>
<td>18,6</td>
<td>18</td>
<td>61376</td>
<td>3,0</td>
<td>Beja e Almodôvar</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>111</td>
<td>99829</td>
<td>4,9</td>
<td>Vidigueira e Moura</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>115</td>
<td>7258</td>
<td>0,4</td>
<td>Mértola</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>122</td>
<td>104130</td>
<td>5,2</td>
<td>Mértola</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>124</td>
<td>8053</td>
<td>0,4</td>
<td>Mértola</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>126</td>
<td>69653</td>
<td>3,5</td>
<td>Moura</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>127</td>
<td>24386</td>
<td>1,2</td>
<td>Mértola</td>
</tr>
<tr>
<td></td>
<td>670140</td>
<td>33,2</td>
<td>13</td>
<td>37792</td>
<td>1,9</td>
<td>Montemor-o-Novo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>110</td>
<td>66176</td>
<td>3,3</td>
<td>Serpa</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>120</td>
<td>49204</td>
<td>2,4</td>
<td>Arronches e Portalegre</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>121</td>
<td>319293</td>
<td>15,8</td>
<td>Évora, Montemor, Arraiolos e Aandroal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>123</td>
<td>12437</td>
<td>0,6</td>
<td>Serpa</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>128</td>
<td>49916</td>
<td>0,6</td>
<td>Almodôvar</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>129</td>
<td>135322</td>
<td>6,7</td>
<td>Elvas, Fronteira e Monforte</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>99570</td>
<td>4,9</td>
<td>Portel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>117</td>
<td>3613</td>
<td>0,2</td>
<td>Elvas</td>
</tr>
<tr>
<td></td>
<td>224850</td>
<td>11,1</td>
<td>130</td>
<td>121667</td>
<td>6,0</td>
<td>Estremoz e Sousel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>34367</td>
<td>1,7</td>
<td>Odemira</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>27583</td>
<td>1,4</td>
<td>Alcácere do Sal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>7161</td>
<td>0,4</td>
<td>Avis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>73681</td>
<td>3,6</td>
<td>Ponte de Sôr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>176744</td>
<td>8,8</td>
<td>Alcácere do Sal e Grândola</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>113</td>
<td>9743</td>
<td>0,5</td>
<td>Odemira</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>114</td>
<td>5784</td>
<td>0,3</td>
<td>Almodôvar</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>118</td>
<td>145323</td>
<td>7,2</td>
<td>Santiago do Cacém e Odemira</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>119</td>
<td>2400</td>
<td>0,1</td>
<td>Santiago do Cacém</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>125</td>
<td>38153</td>
<td>1,9</td>
<td>Alcácere do Sal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>131</td>
<td>38231</td>
<td>1,9</td>
<td>Avis</td>
</tr>
<tr>
<td></td>
<td>559269</td>
<td>27,7</td>
<td>112</td>
<td>56407</td>
<td>2,8</td>
<td>Crato e Gavião</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>116</td>
<td>75880</td>
<td>3,8</td>
<td>Castelo de Vide e Nisa</td>
</tr>
<tr>
<td></td>
<td>132286</td>
<td>6,6</td>
<td>14</td>
<td>57690</td>
<td>2,9</td>
<td>Beja e Ferreira do Alentejo</td>
</tr>
</tbody>
</table>

Source: Project team, based on K-means analysis