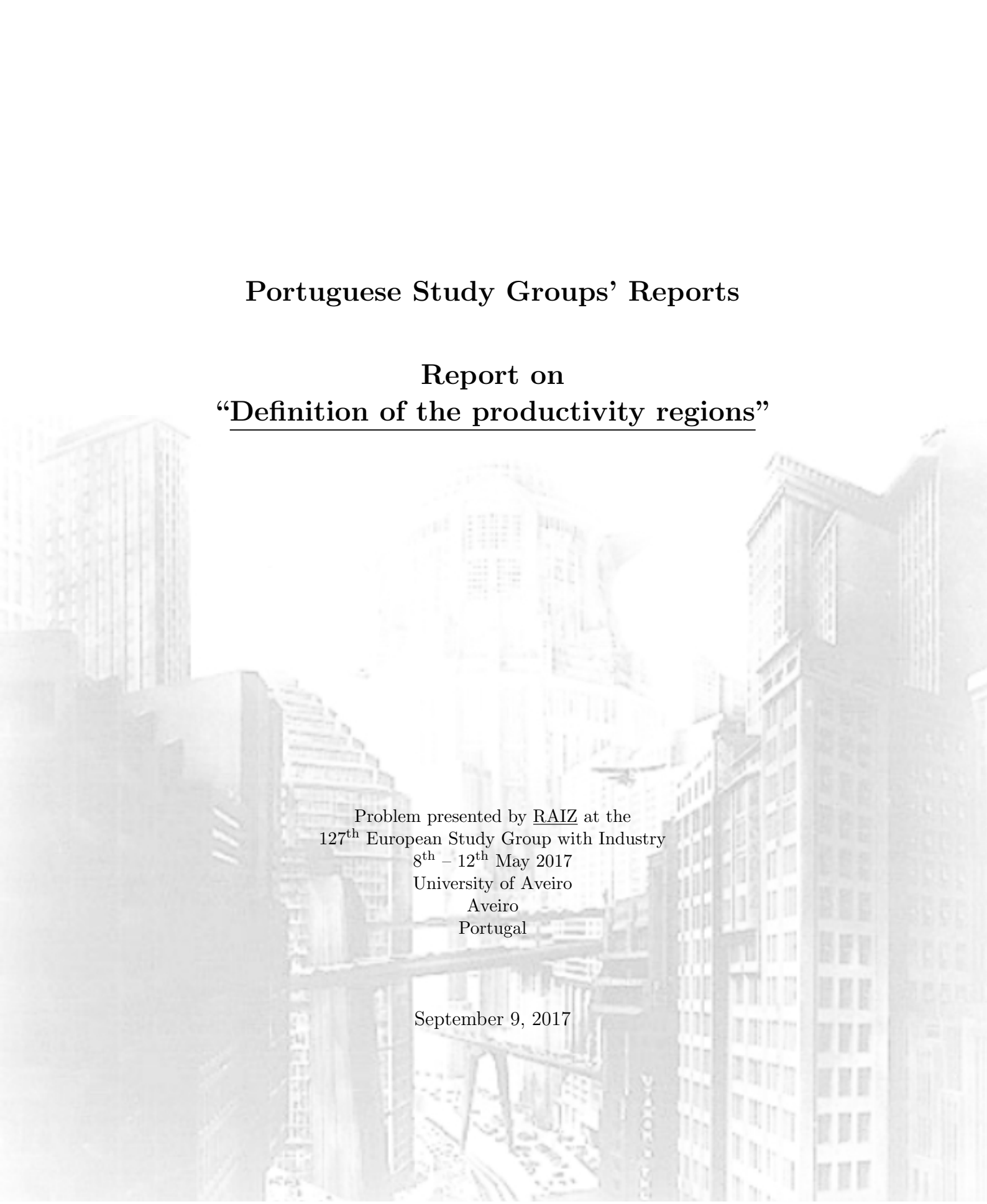

Portuguese Study Groups' Reports

Report on "Definition of the productivity regions"

Problem presented by RAIZ at the
127th European Study Group with Industry
8th – 12th May 2017
University of Aveiro
Aveiro
Portugal

September 9, 2017

Problem presented by: Catarina Silva and Isabel Pinto (RAIZ)

Study group contributors: Adérito Araújo, Alexandra Gavina, Ana Tavares, J. Orestes Cerdeira, Jorge Santos, Manuel Cruz and Sílvia Barbeiro

Report prepared by: J. Orestes Cerdeira (e-mail: jo.cerdeira@fct.unl.pt)
Manuel Cruz (e-mail: mbc@isep.ipp.pt)
and
Alexandra Gavina (e-mail: alg@isep.ipp.pt)

Executive summary

Eucalyptus productivity is strongly related with climate and soil types of the area where it is planted. To accurately assess the potential productivity of that species in Portugal, plantations were monitored at different management units compartments (MUC) at several locations all over Portugal. Certain indices of productivity of the Eucalyptus at each MUC were recorded, as well as the type of climate and soil characteristics of the region. Both climate and soil, factors that affects Eucalyptus grow, were classified in in ten classes $1, 2, \ldots, 10$ of expected growing productivity for the Eucalyptus. Thus, every MUC belongs to a unique pair (c, s), with $1 \leq c \leq 10$ and $1 \leq s \leq 10$ indicating the type of climate and the type of soil of the region where MUC is located, respectively, and it is expected that to have high (low) productivity indices when c and s are both close to 10 (1).

The aim of this work is to identify regions that have similar productivity levels based on the classifications of soil and climate types and to check if the available data provided by RAIZ show that those factors affect the productivity indices.

During the 5-days ESGI this team worked on the datasets provided by RAIZ, presented an update for the existing MAI productivity chart and developed new clusters for the Density, Yeld and Consumption productivity indices. The definition of some quality measures for the clusters, allowed to compare the different approaches and also point out some fragilities on the datasets. Indeed, a review of classification regarding climate and/or soil characteristics is suggested, as well as the need of a bigger sample for the Density, Yeld and Consumption productivity indices in order to get more reliable outputs.

1 The Challenge

RAIZ Forest and Paper Research Institute is a private, non-profit organization whose objective is to strengthen the competitiveness of the forestry (in particular the eucalyptus wood), in order to maximize access and quality of the eucalyptus wood for pulp and paper production.

Eucalyptus productivity is strongly related with the climate and soil types of the plantation plots and it has been subject of several studies (see, for instance, [1] or [5]). To analyze the effect of these factors on productivity, climate and soil were independently classified by RAIZ in ten classes $1, 2, \ldots, 10$ of expected growing productivity for that species, and sites all over Portugal were assigned to pairs (c, s), with $1 \leq c \leq 10$ and $1 \leq s \leq 10$, indicating the type of climate and the type of soil of the corresponding region. High (low) productivity is expected to occur in locations where both c and s are close to 10 (1).

The challenge proposed by RAIZ on this study group consists on identifying regions, possibly with different (c, s) values, that have similar productivity levels.

To this purpose plantations were monitored at different management units compartments (MUC) at several locations all over Portugal. Four indices were considered to estimate the productivity of Eucalyptus. A value of each of these indices was calculated at each MUC (see Section 2 below), and the pair (c, s) quantifying the climate and soil types of the region where MUC is located was also recorded. Thus, we have for each pair (c, s) $(1 \leq c, s \leq 10)$ and, for each of the four productivity indices, n_{cs} values λ quantifying the productivity at each of the n_{cs} MUC that are located in regions classified as (c, s) in terms of their climate and soil types, respectively. It should be noted that for some pairs (c, s), $n_{(c,s)} = 0$, i.e., no MUC belongs to regions classified as (c, s). We denote by G the set of pairs (c, s) for which $n_{(c,s)} > 0$. We also denote by Λ_{cs} the collection of the n_{cs} values λ recorded at MUC on regions classified (c, s), and we let $\Lambda = \cup_{(c,s)\in G}\Lambda_{cs}$ be the collection of all values w.r.t. a given productivity index. Finally we let n denote all the values in Λ.

The challenge is to determine partitions of G so that climate and soil types assigned to the same cluster have similar productivity, and climate and soil types assigned to different clusters have distinct productivity levels.

RAIZ advanced as benchmark an 8-partition of G which we will refer as the RAIZ-partition.

We give an approach for finding partitions of G, propose two indicators to estimate the quality of the partitions, and report and discuss results on these proposals, for different numbers of clusters of the partitions, using the data provided by RAIZ.

2 The Data

The datasets provided by RAIZ, consisted of 4 productivity outputs:

- Mean Annual Increment (MAI);
- Density;
- Yeld/Effiency (Yeld); and
- Consumption.

MAI values were recorded in a dataset for each of the 32547 MUC presented on the sample. The other three outputs, were registered in a separated file with respect to 897 MUC.

In this section we present some descriptive statistics to have some insight on each of these datasets, in order to better understand the results presented in Section 4. Those samples main descriptive measures are summarized on Table 1, where the last column (C.V.) represents the coefficient of variation (i.e. the ratio of the standard deviation to the mean).

Dataset	n	min	1^{st} Qt	median	3^{rd} Qt	max	mean	std	C.V.
MAI	32547	0.9	4.8	7.2	10.8	33.8	8.4	4.9	0.59
Density	897	425.3	530.8	563.7	598.1	722.8	563.6	48.7	0.09
Yeld	897	34.5	48.9	51.1	53.5	61.8	51.0	3.5	0.07
Consumption	897	2.3	2.9	3.1	3.4	4.9	3.2	0.4	0.12

Table 1: Main descriptive measures for each of the four productivity outputs.

2.1 MAI dataset

In this dataset were considered eucalyptus from 8 to 13 years old, with 1 to 4 stand rotations[1], resulting in 32547 samples grouped according to climate and soil types (c, s), with $c, s = 1, \ldots, 10$. The number of cells (c, s) with at least one MAI value, i.e., the size of G is 72.

For coherence with the format of data presented by RAIZ representatives at the ESGI127, we will represent climate type on yy's axis while soil type will be represented with a reversed scale on xx's axis.

In Figure 1 we can see the data grouped in 10×10 cell grid where, in each cell (c, s), it is indicated the number of MAI values on that cell. The colors

[1]The rotation length is a key component of even-aged forest management systems

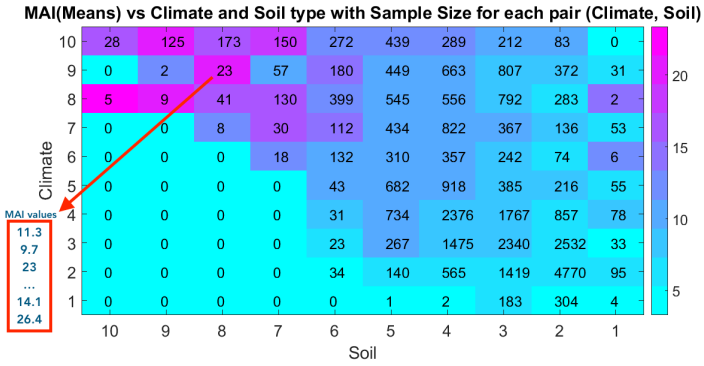

Figure 1: MAI: Number of cases observed for each cell.

are indexed by the mean MAI values on cells. The color scale is depicted on the right-hand side of the plot. The box on bottom-left of Figure 1 details some of the 23 MAI-values belonging to a particular cell $(9, 8)$.

In Figure 2, we present a color map for four central tendency measures. The mean is represented on top left, the median on top right plot, while the 1^{st} (3^{rd}) quartile is represented on left (right) bottom plot.

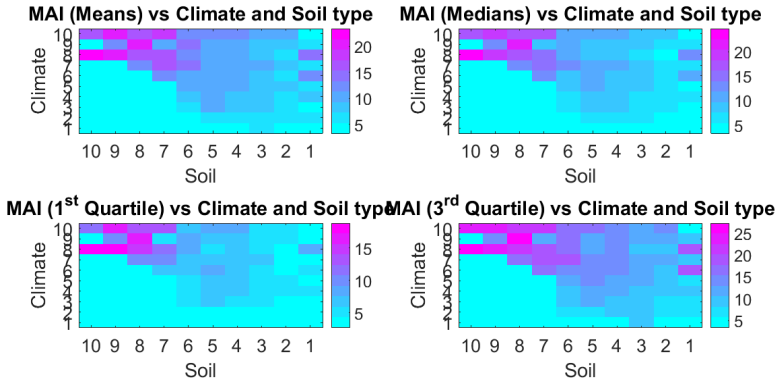

Figure 2: MAI: Mean and percentiles for each cell.

In order to provide a deeper insight into the distribution of MAI values, Figure 3 presents the MAI empirical cumulative distribution, the histogram as well as the data box-plot. We also present the Komolgorov-Smirnof (KS) p-value result for the normality test. Normality is clearly rejected by KS-test for any usual significance level and there is a considerable number of MAI extreme values as represented on the box-plot.

The mean of MAI was computed considering fixed one of the coordinates of the pair (c, s) and varying the other. In Figure 4 we represent the mean of MAI for fixed c and Figure 5 show us the mean MAI for fixed s.

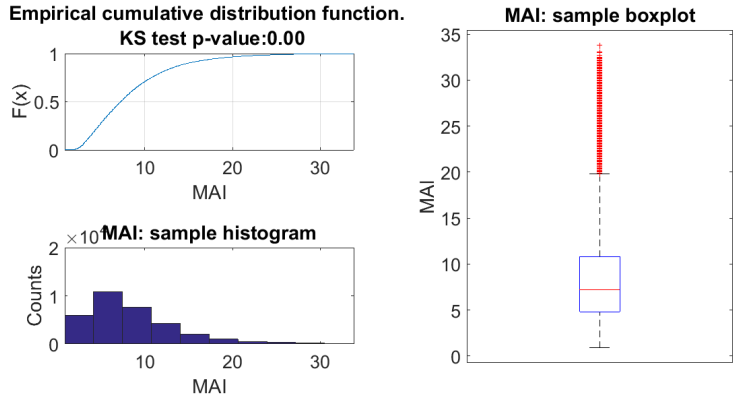

Figure 3: MAI: Sample distribution overview.

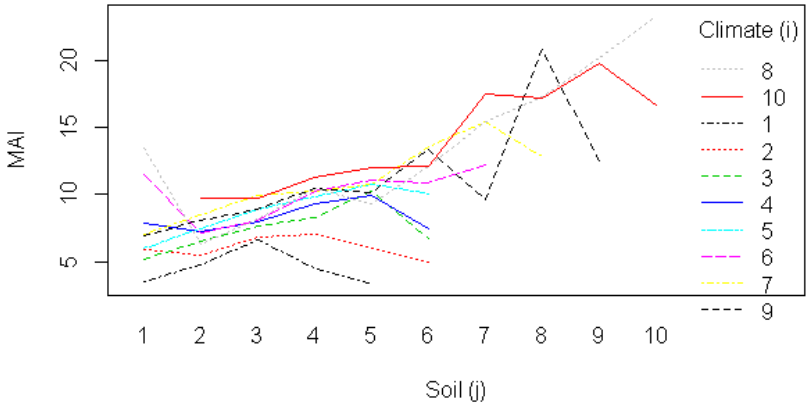

Figure 4: Mean of MAI with fixed climate (c).

From the observations on the plots represented in figures 4–5, it seems to be a positive correlation from MAI with soil type, and even with climate.

However, the Box-plots for the pairs (Climate, MAI) and (Soil, MAI) presented in Figure 6 also highlighted the high variability of MAI values, specially evident for low classifications on Soil type, and for almost all the values of Climate.

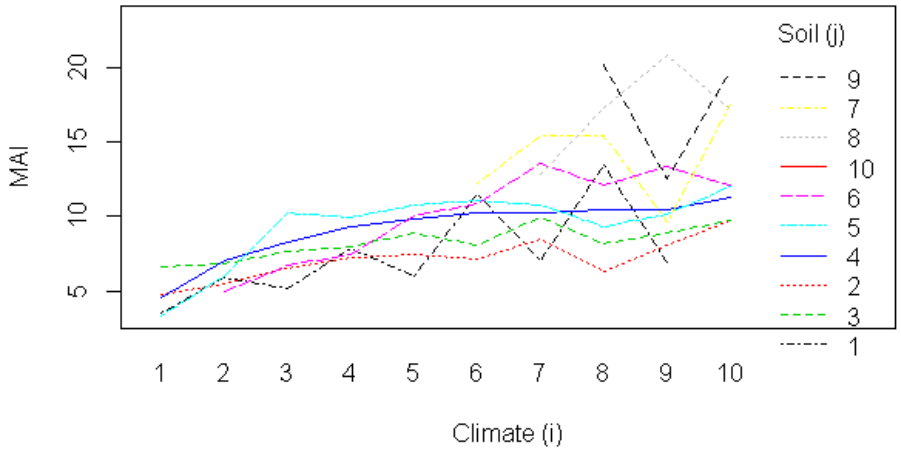

Figure 5: Mean of MAI with fixed soil (s).

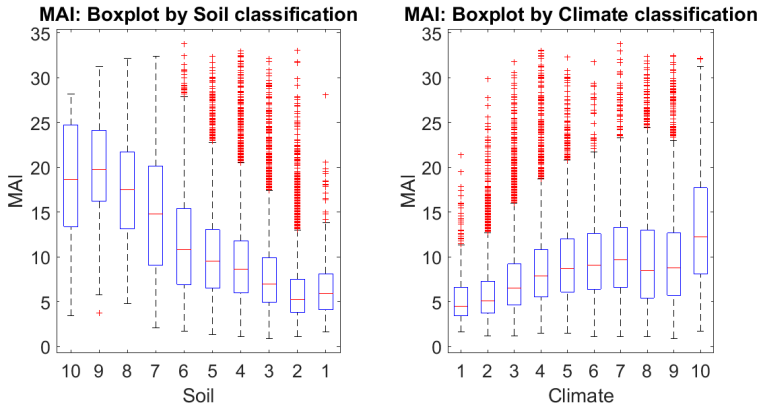

Figure 6: MAI: Soil and Climate box-plots.

2.2 Density, Yeld and Consumption dataset

This analysis proceed now with a closer look at the second dataset provided by RAIZ, that include the values for Density, Yeld and Consumptions instances, and consists on a sample of size 897, arising from 33 different combinations on (c, s), where the climate indexes varies from 1 to 10, and soil from 1 to 8.

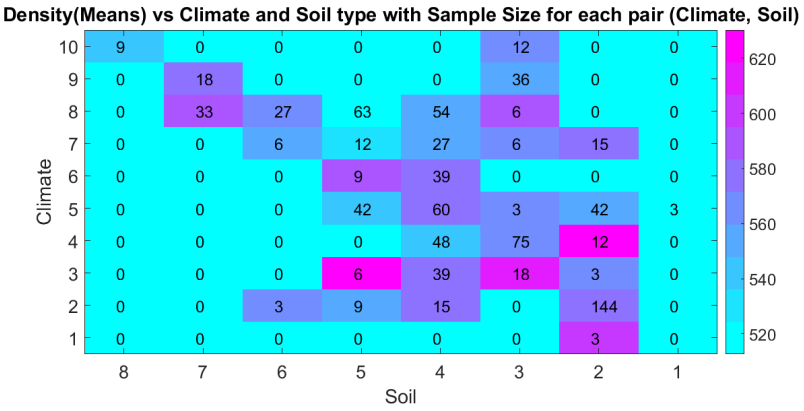

Figure 7: Density: Mean and Sample Size for each cell.

2.2.1 Density

The Density instance represents the continuous variable, Basic Density, that is measured in (kg/m^3). According to [4] basic density is considered the main indicator of wood quality, as it correlates with all other wood properties, including retractability, mechanical properties and anatomy. According to the same source, density affects all processes in which wood is present, including pulping, charring, machining and log breakdown.

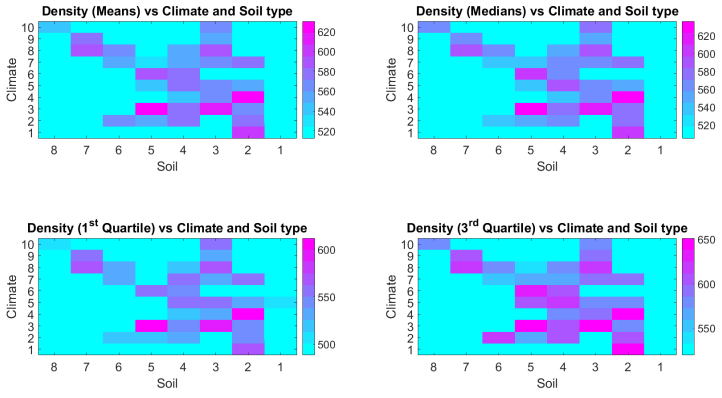

Figure 8: Density: Mean and percentiles for each cell.

The distribution of Density values on the sample are reasonable symmetric around the mean, having small variation with respect to sample mean (C.V. ≈ 0.09). The box-plot presented on Figure 10 doesn't seem to show any relevant trend w.r.t. soil type.

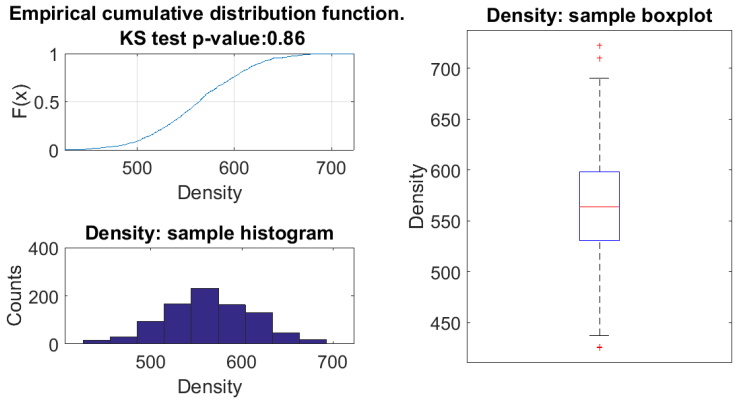

Figure 9: Density: Sample distribution.

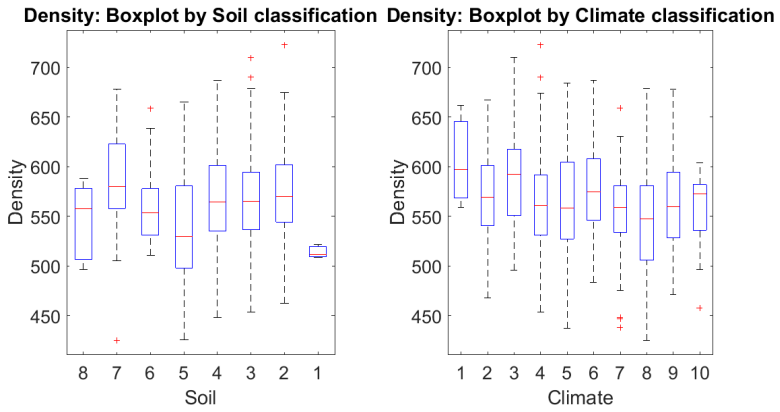

Figure 10: Density: Soil and Climate box-plots.

2.2.2 Yeld

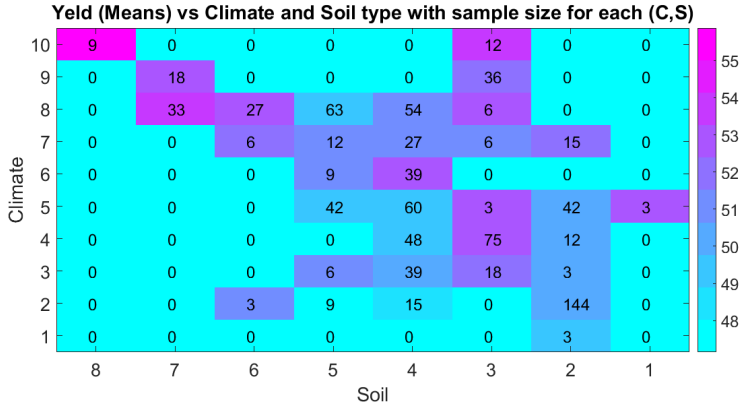

Figure 11: Yeld: Mean and Sample Size for each cell

The Yeld/Efficience is an non-dimensional continuous variable, with 897 observations presented in the sample, concentrated in a narrow band of values and which coefficient of variation is very small (around 7%).

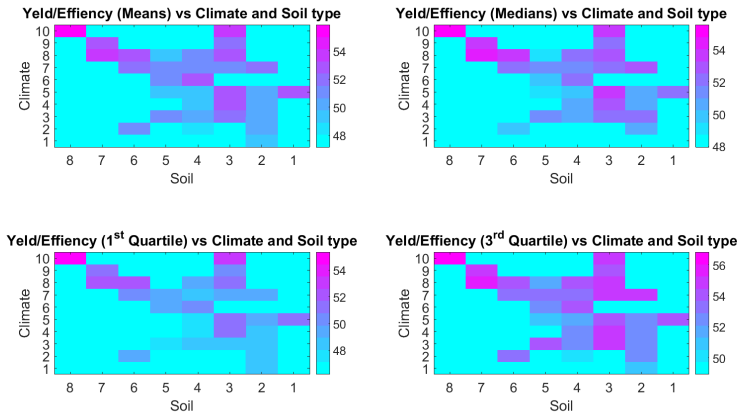

Figure 12: Yeld: Mean and percentiles for each cell.

The box-plot presented on Figure 14 seems to show a small increase of this parameter with the climate classification, but is rather inconclusive regarding the soil classification.

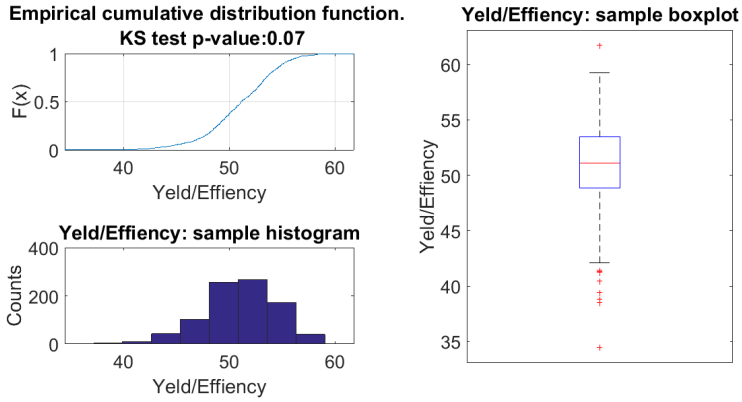

Figure 13: Yeld: Sample distribution.

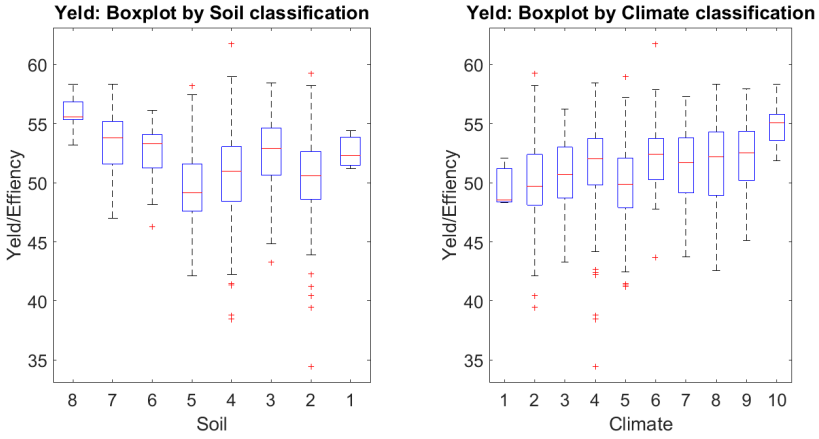

Figure 14: Yeld: Soil and Climate box-plots.

2.2.3 Consumption

The instance Consumption represents the specific consumption measured in (m^3/tAD), which is a continuous variable that is calculated using Basic Density and Yeld, through the formula $Consumption = \frac{900}{\frac{Density * Yeld}{100}}$.

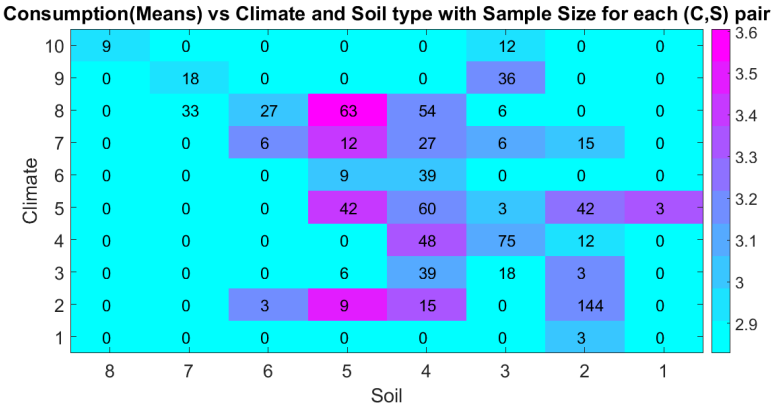

Figure 15: Consumption: Mean and Sample Size for each cell.

The distribution of Consumption values in the sample is quite asymmetric, and the K-S test rejects the null hypotheses for any usual level of significance. There is a significant number of observations which Consumption is greater than the upper whisker of $Q_3 + 1.5(Q_3 - Q_1)$.

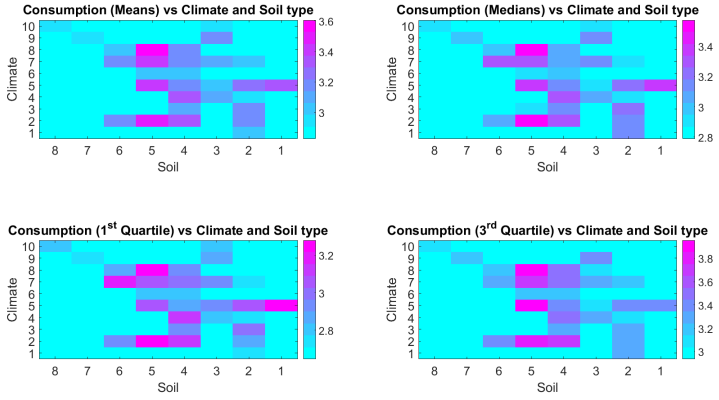

Figure 16: Consumption: Mean and percentiles for each cell.

Finally, the observation of the several representations of the data, don't show any evident relation between consumption and the two Soil and Climate classifications.

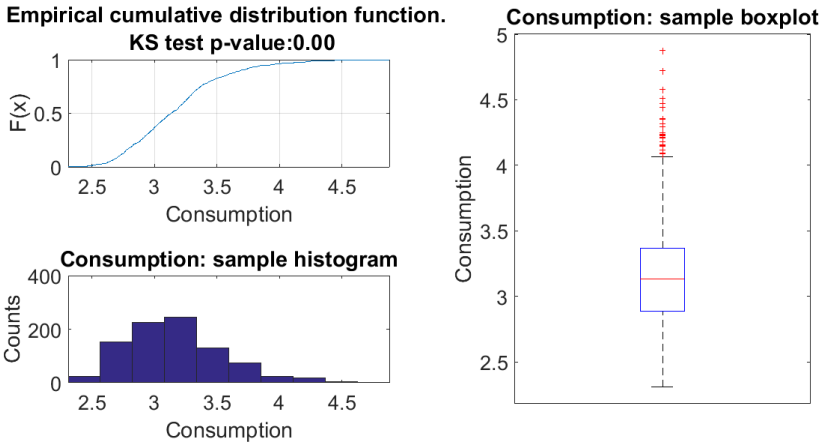

Figure 17: Consumption: Sample distribution.

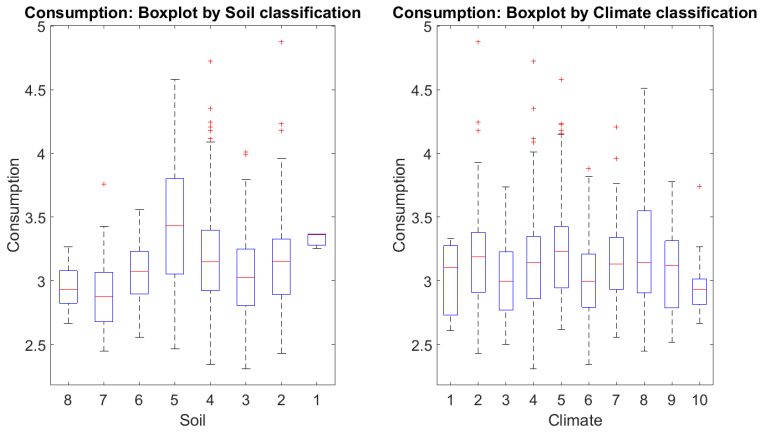

Figure 18: Consumption: Soil and Climate box-plots.

3 Methods

In this section we give an algorithm for finding partitions of the cells of G, aiming that cells clustered together have similar productivity levels, and cells on different clusters have distinct productivity. We also propose three indicators to assess the quality of the produced clusters w.r.t . productivity.

3.1 Determining partitions of G

We propose the following two step procedure to determine partitions of G w.r.t. each productivity index.

Procedure *Partition*:

step 1 Define a k-partition of Λ, consisting of intervals $[a_{i-1}, a_i], i = 1, \ldots, k$ with $a_0 = \min\{\Lambda\}$ and $a_k = \max\{\Lambda\}$.

step 2 For every $(c, s) \in G$, calculate the value of a centrality measure D_{cs} of the points in Λ_{cs}, and assign color i to cells (c, s) for which $D_{cs} \in [a_{i-1}, a_i]$.

In our computational experiments, to find the partitions on step 1, we used the clustering methods k-means [2], k-medoids [6] and kernel density estimation (KDE) [3], and on step 2 we used as centrality measures the mean and the median, i.e., $D_{cs} = \text{Mean}(\Lambda_{\text{cs}})$ or $D_{cs} = \text{Median}(\Lambda_{\text{cs}})$.

Figure 19 shows the results of procedure *Partition* working on RAIZ data with productivity MAI, for $k = 6$, using the cluster method k-means to obtain the 6-partition of Λ consisting of the 32547 MAI values (Figure 19 (top)), and letting the centrality metric $D_{cs} = \text{Mean}(\Lambda_{\text{cs}})$. The resulting 6-partition of G is shown in Figure 19 (bottom).

Note that the number of clusters of the partition of G obtained at step 2 may be less than k, the number of clusters at step 1 (this happens if some interval $[a_{i-1}, a_i]$ includes no D_{cs}).

3.2 Assessing the quality of partitions of G

Given a partition of G with colors $i = 1, \ldots, k$, we denote by $i(c, s)$ the color of cell (c, s) on that partition.

To estimate the Eucalyptus productivity on cells with color i a possibility could be

$$E1(i) = D(\cup_{\{(c,s):j(c,s)=i\}} \Lambda_{cs}) \tag{1}$$

i.e. a centrality value of the set of all points in the cells of G with color i in the partition of G.

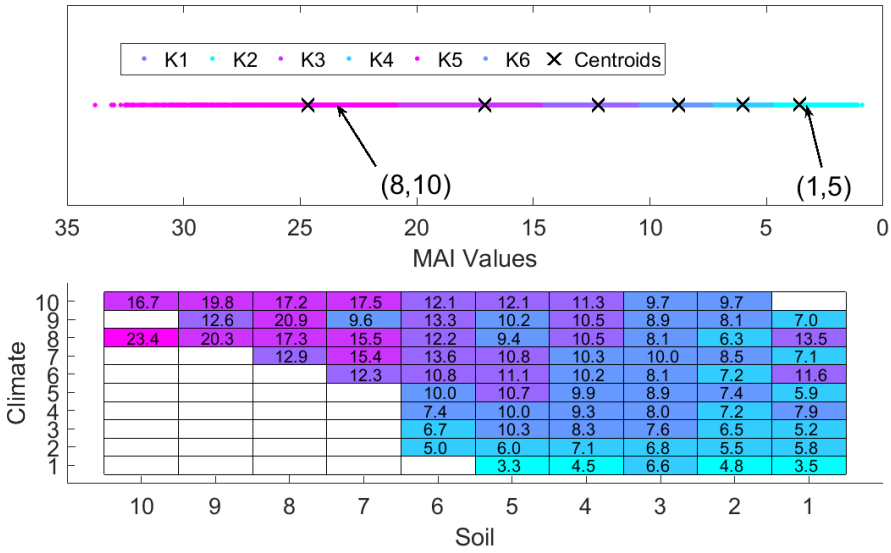

Figure 19: An illustration of procedure *Partition* with input MAI dataset. Top: 6-partition obtained at step 1 using k-means method. Bottom: the corresponding partition of G obtained at step 2 using centrality $D_{cs} = \text{Mean}(\Lambda_{cs})$.

Another possibility (which does not depend on the partition of G) could be

$$E2(i) = D(\Lambda \cap [a_{i-1}, a_i]) \tag{2}$$

i.e., a centrality estimate of the points of Λ that at step 1 of the procedure *Partition* were clustered together in interval $[a_{i-1}, a_i]$.

Thus, if in a partition of G a point λ is in a cell having color i, $|\lambda - E1(i)|$ and $|\lambda - E2(i)|$ are estimates of the dissimilarity of λ w.r.t. to the reference values for color i, $E1(i)$ and $E2(i)$, respectively.

Hence, we devise the following formula to assess the quality of k-partition of G:

$$Ej = \frac{1}{n} \sum_{(c,s)\in G} \sum_{\lambda \in \Lambda_{(c,s)}} \frac{|\lambda - Ej(i(c,s))|}{\lambda} \tag{3}$$

where $n = |\Lambda|$ is the total number of observations, and $j = 1, 2$ specifies which criterion (1) or (2) is used.

In addition to expression (3), to evaluate the quality of a partition, we also calculate the number of disagreements between the partitions obtained by procedure *Partition* at steps 1 and 2. More specifically, we count the proportion of $\lambda \in \Lambda$ that received different colors in the two partitions. We denote by Mismatches the proportion of these λ.

4 Results and discussion

We applied the procedure *Partition* of Section 3, using different clustering methods at step 1, to each of the four productivity outputs. When the k-medoids method was used at step 1, the median was used as the centrality metric at step 2, and also to compute $E1$ and $E2$. When KDE and k-means were used at step 1, the centrality mean was applied at step 2, and to calculate $E1$ and $E2$.

4.1 MAI dataset

For MAI dataset the main results are presented on Table 2

Method	k step 1	# clusters step 2	$E1$	$E2$	PMismatches
KDE	9	9	0.5048	0.4935	0.8467
k-medoids	9	9	0.5084	0.4525	0.7849
k-means	9	8	0.5071	0.5044	0.7992
Raiz solution	8	8	0.5151	0.6150	NA
KDE	8	8	0.5052	0.5047	0.8302
k-medoids	8	8	0.5161	0.4453	0.7484
k-means	8	7	0.5092	0.4933	0.7558
KDE	7	7	0.5098	0.5013	0.8091
k-medoids	7	7	0.5106	0.4558	0.7017
k-means	7	6	0.5162	0.5061	0.7316
KDE	6	6	0.5101	0.5067	0.7763
k-medoids	6	6	0.5126	0.4616	0.6412
k-means	6	6	0.5169	0.5258	0.6998
KDE	5	5	0.5140	0.5053	0.7212
k-medoids	5	5	0.5215	0.4621	0.5666
k-means	5	5	0.5279	0.5650	0.6710

Table 2: MAI: Tests results.

A remark that follows promptly from the results of Table 2 is that the estimates of the quality of the partitions are large indicating poor separability among the clusters of cells of G. Especially, the values in column

PMismatches which are the proportions of the number of MAI values that received different colors on the two partitions (step 1 and step 2 of procedure *Partition*), are quite high (most values around 0.7 or 0.8), and increase with the number of clusters considered. This fact is a consequence of the great variability of MAI values in each cell (c, s), as noted already in Section 2. See, e.g., in Figure 20 the distribution of the 822 MAI values that are in cell (7,4). In that figure we also represented the centrality measure $D_{cs} = Mean(7, 4)$ (dotted blue line), and the cluster borders in the partition obtained by step 1, using k-means (red lines). About 70% of the MAI values of that cell lay outside the cluster borders. Moreover, the range of MAI values of this cell almost equals the whole interval of variation of all MAI values (see Figure 3 and Table 1).

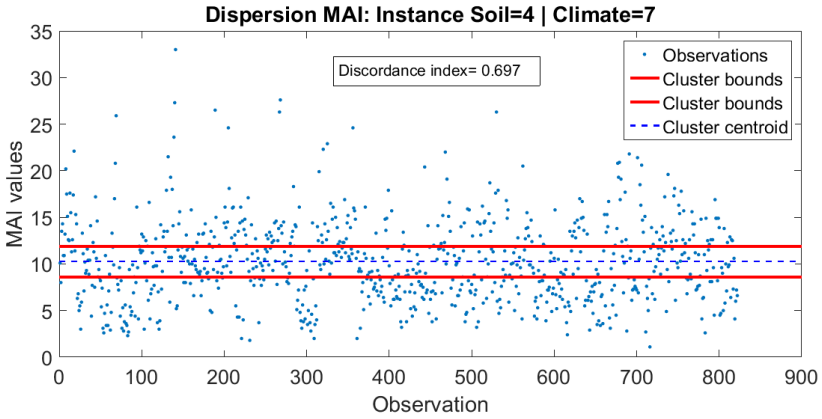

Figure 20: MAI dispersion values within cell (7,4).

To explain the large values of $E1$, we calculated $E1$ for the (trivial) 72-partition of G where every cluster consists of a single cell and obtained $E1^* = 0.5022$. Note that this is the "best" partition of G one can get. Hence the large value of $E1^*$ is a clear indication of a poor correlation between MAI vales and the classification established for climate and soil types. This clear explains the high values of column $E1$, and allows to conclude that clusters obtained with our approach, e.g., KDE method with $k = 8$ or KDE with $k = 7$ (which only add to $E1^* = 0.5022$, which is the minimum possible $E1$ value, 0.003 and 0.0076, respectively) are quite reasonable taking into account the distribution of MAI values amid cells.

The same observations apply to indicator $E2$. It is worth noting that RAIZ solution presented the worst $E2$ value among the solutions in Table 2. We display in Figure 22 the solution that attained the lowest value of $E2$ (see Table 2). That solution, obtained with k-Medoids consisting of eight clusters, and that in Figure 21, also with $K = 8$, obtained using KDE, which has the lowest $E1$ value, seem plausible options. If seven clusters were desirable then KDE with $k = 7$ that only adds 0.0076 to $E1^* = 0.5022$

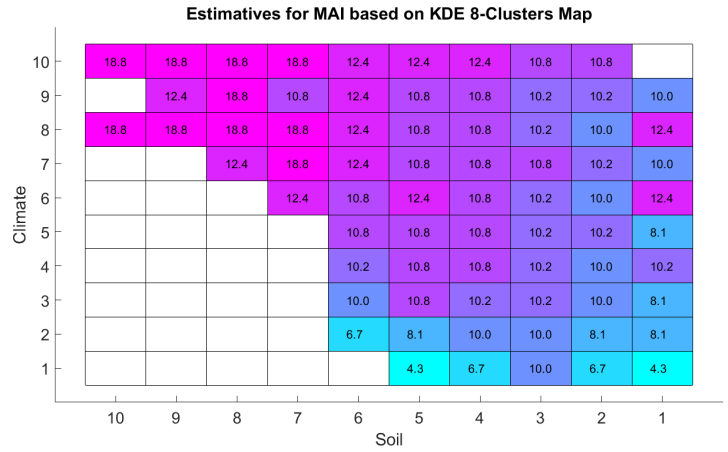

Figure 21: Partition obtained for MAI, using KDE method with $k = 8$.($E1$=0.5052)

could be a good choice.

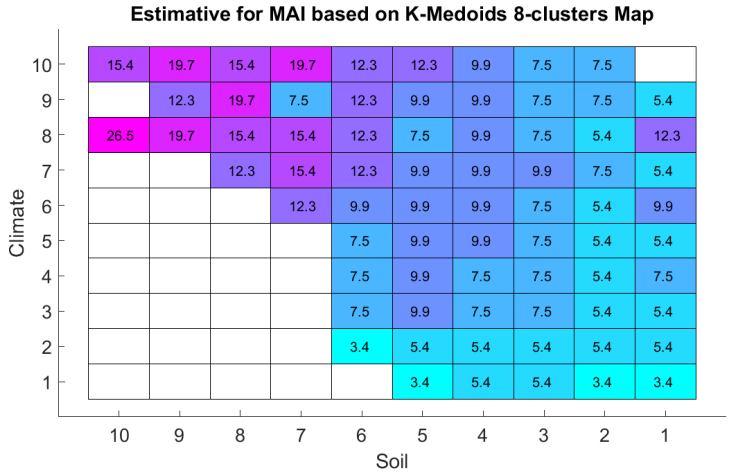

Figure 22: Partition obtained for MAI, using K-medoids method with $k = 8$. ($E2$=0.4453)

4.2 Density, Yeld and Consumption dataset

We performed a similar analysis for the other productivity outputs. However, the reader should be aware that in this dataset there are only 33 cells in G (i.e., cells with at least one value), and only 21 cells with 10 or more observations (Figure 7). Therefore, results should be cautiously considered.

We present in Table 3 the solution and corresponding $E1$, $E2$ and PMis-

Productivity	Method	k	$E1^*$	$E1$	$E2$	PMismatches
Density	KDE	6	0.0616	0.0621	0.0867	0.76
Yeld	KDE	6	0.0481	0.049	0.0486	0.78
Consumption	KDE	5	0.0849	0.0860	0.62	0.77

Table 3: Density, Yeld and Consumption: Tests results.

matches values, that for each productivity indicator, satisfied the following conditions. The number of clusters k was selected as the minimum for which $E1 - E1^* < 0.01$. This was achieved for $k = 6$, for Density and Yeld, and $k = 5$, for Consumption. Next we selected the solution with the lowest $E2$. For all productivity indices this was obtained with the KDE clustering method. The solutions are displayed in Figures 23, 24 and 25.

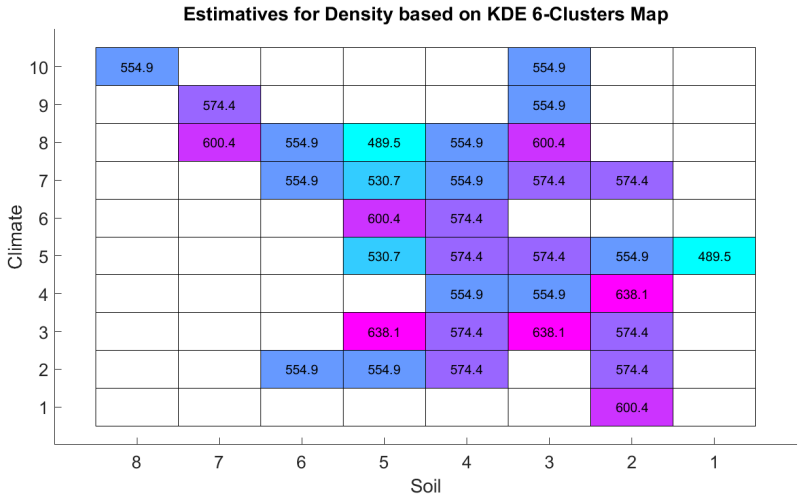

Figure 23: Proposed cluster for Density, with estimate value for each (c,s) pair.

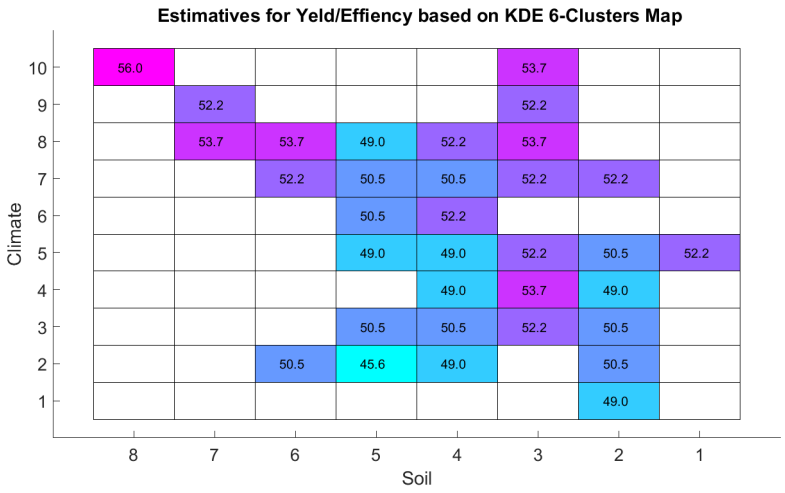

Figure 24: Proposed cluster for Yeld, with estimate value for each (c,s) pair.

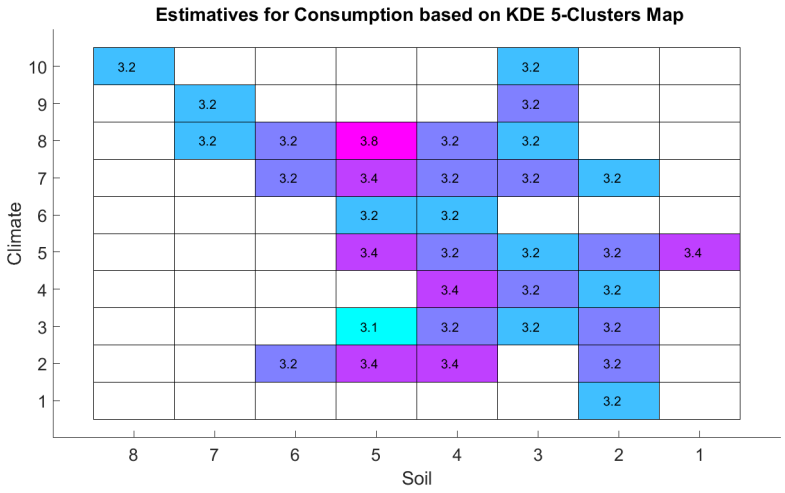

Figure 25: Proposed cluster for Consumption, with estimate value for each (c,s) pair.

5 Conclusions and recommendations

Comments, remarks and recommendations from this preliminary study follows.

- Concerning MAI.

  - It seems there exists a positive correlation between the distribution of MAI values within cells and both soil and climate types.
  - Nevertheless, there is a significant variability of MAI values in the cells.
  - The high values of $E1^*$ is a clear indication that the proposed attempt for the classification of climate and soil characteristics with regard to Eucalyptus productivity should be reviewed.

- Concerning Density, Yeld and Consumption.
  - With respect to Density, no correlation is apparent between distribution of productivity values in cells and climate and soil classifications. This follows from data analysis on Section 2, and from the distribution of colors on the 6-partition of Figure 23.
  - With respect to Yeld, the 6-partition of Figure 24 indicates a trend of growing productivity reference values in cells with the increase type of climate. This is in accordance with data description of Section 2.
  - With respect to Consumption, we found no evidence of a correlation between productivity and climate and soil types (see the boxplot in Figure 18, and the pattern of colors of the 6-partition provided in Figure 25).
  - But the main drawback regarding the analysis on this dataset is a clear lack of information on Eucalyptus productivity on the 10×10 grid cells of different climate and soil type, and on the number of values within cells. Reliable results for such analysis require overcoming this issue with the addition of more information.

References

[1] Almeida, Auro C and Siggins, Anders and Batista, Thiago R and Beadle, Chris and Fonseca, Sebastião and Loos, Rodolfo, *Mapping the effect of spatial and temporal variation in climate and soils on Eucalyptus plantation production with 3-PG, a process-based growth model*, Forest Ecology and Management, vol. 259, 9, pp. 1730–1740, 2010, Elsevier

[2] Arthur, David and Vassilvitskii, Sergei *K-means++: The Advantages of Careful Seeding*, Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, 2007, Society for Industrial and Applied Mathematics

[3] Bowman, A. W. and A. Azzalini *Applied Smoothing Techniques for Data Analysis*, 1997, Oxford University Press Inc.

[4] Couto, Allan Motta and Trugilho, Paulo Fernando and Neves, Thiago Andrade, and Protásio, Thiago de Paula, and Sá, Vânia Aparecida *Modeling of basic density of wood from Eucalyptus grandis and Eucalyptus urophylla using nondestructive methods*, CERNE, vol. 19, 1, pp. 27–34, 2012.

[5] Miehle, Peter and Battaglia, Michael and Sands, Peter J. and Forrester, David I. and Feikema, Paul M. and Livesley, Stephen J. and Morris, Jim D. and Arndt, Stefan K. *A comparison of four process-based models and a statistical regression model to predict growth of plantations*, Ecological Modelling, Volume 220, 5, pp. 734–746, 2009, Elsevier.

[6] Hae-Sang, Park and Chi-Hyuck, Jun *A simple and fast algorithm for K-medoids clustering*, Expert Systems with Applications, vol. 36, 2, pp. 3336–3341, 2009, Elsevier.