Reducing the loss of information and gaining accuracy with clustering methods in a global land-use model

Jan Philipp Dietrich[a,b,*], Alexander Popp[a], Hermann Lotze-Campen[a]

[a] *Potsdam Institute of Climate Impact Research (PIK), Telegraphenberg A 31 14473 Potsdam, Germany*
[b] *Department of Physics, Humboldt University, Newtonstr. 15 12489 Berlin, Germany*

Abstract

Global land-use models have to deal with processes on several spatial scales, ranging from the global scale down to the farm level. The increasing complexity of modern land-use models combined with the problem of limited computational resources represents a challenge to modelers. One solution of this problem is to perform spatial aggregation based on a regular grid or administrative units such as countries. Unfortunately this type of aggregation flattens many regional differences and produces a homogenized map of the world. In this paper we present an alternative aggregation approach using clustering methods. Clustering reduces the loss of information due to aggregation by choosing an appropriate aggregation pattern.

We investigate different clustering methods, examining their quality in terms of information conservation. Our results indicate that clustering is always a good choice and preferable compared to grid-based aggregation. Although all the clustering methods we tested delivered a higher degree of information conservation than grid-based aggregation, the choice of clustering method is not arbitrary. Comparing outputs of a model fed with original data and a model fed with aggregated data, bottom-up clustering delivered the best results for the whole range of numbers of clusters tested.

Keywords: aggregation, downscaling, clustering, information conservation, agricultural model, land use model, scale

1. Introduction

An important step in the analysis of a process is to split it into sub-processes or sub-objects. One very helpful approach is to classify processes or objects based on their scale. According to Turner et al. [30, p. 27] "scale refers to the spatial or temporal dimension of an object or process". In the case of objects it is related to their size or life-time, in the case of processes it is related to a characteristic time span or spatial extent, for instance the duration or length of a periodic process.

To describe a range of several scales, for instance to characterize the range of scales covered by a model, two further terms are used: grain and extent [9, 30, 17]. Grain (often also called "resolution") is the smallest unit (temporal or spatial) of a data set, model, or an observation, for instance the grid size of a spatially explicit data set. Extent describes the total spatial or temporal coverage and is the upper scale limit.

Effective resolution is the precision or level of detail of a measurement. Often grain is already a good indicator for effective resolution. However, in some situations this relation does not hold. Examples are disturbances in a measurement, so that differences between adjacent cells are masked by the noise. In that case a further decrease in grain size does not deliver additional detail and does not lead to an increased effective resolution. Grain size is always related to the physical characteristics of a data set (size of a single

*Corresponding author. Tel.: +49 331 288 2440, Email address: dietrich@pik-potsdam.de
Potsdam Institute for Climate Impact Research, P.O. Box 60 12 03, 14412 Potsdam, Germany

grid cell), whereas effective resolution refers to the quality of a data set (the precision or detail in which the original system is reproduced by the data set).

The full separation within a model of processes at different scales is often not possible because of cross-scale interactions connecting these processes. Cross-scale interactions play an important role in global change research [35, 6, 11] for several reasons. First, the integration of models and data from different disciplines, such as physics, biology, geography or economics, is typically connected to the issue of different spatial and temporal scales [35]. Second, because of nonlinearities a proper treatment of cross-scale interactions is often a requirement for accurate simulations [6, 11]. Third, the interactions itself are of great interest to understand the dynamics and to be able to assess the impact of policies at different scales [6, 7].

For agricultural land-use models especially the first two issues are highly relevant: One characteristic of land-use models is that they link elements from geography and economics. Since the general approaches of both disciplines differ significantly several scale related problems arise. In geography spatial information plays a major role; data is linked to a location and spatial explicitness is most important. In economics markets and market equilibria are the dominant processes. Spatial patterns are typically neglected in a first order approximation of a system. Instead the focus lies on complex market dynamics and flows of inputs and outputs.

The challenge for agricultural land-use models is to take the dominant aspects of both domains into account: global markets and their market equilibria and spatially varying environmental conditions and production patterns. However, including high-resolution data into an equilibrium model leads to significant computational problems. Increasing the number of simulated units typically leads to a disproportionate increase in computation time and the required amount of working memory. For instance, the nonlinear land-use model MAgPIE ("Model of Agricultural Production and its Impact on the Environment") [19, 20, 21] shows quadratic increases in computation time with increasing number of simulated cells. So halving the grain side length, which means a quadruplication of 2D-cells, leads to a 16-times longer computation time. Furthermore, the increase in working memory requirements limits the total number of cells to less than 5000.

In current agricultural research several approaches are used to deal with this problem. Models focused on the economy often cover global agricultural markets, but only at a coarse spatial resolution of a few world regions (e.g. AgLU [25], FASOM [2], IMPACT [23]), whereas models focused on geographical or ecological processes either only model certain regions of the world, with exogenous global markets (e.g. CLUE [31, 32, 34], SALU [29]), or apply a rule-based approach (e.g. LandSHIFT [26] - a general land use model review was done by Heistermann et al. [13]). Hence, either the economic or the ecological part is represented in a simplified manner concentrating the formulation of the model either on the global or local scale. A promising but complex approach to overcome these limitations is to couple models that focus on different scales and sectors (e.g. Verburg et al. [33]). Another possibility to cope with this issue is the use of cluster algorithms to increase the effective spatial resolution under a constant number of simulation units (see for example Letourneau et al. [18]). Here we present and compare a selection of clustering algorithms and analyze the benefits of these clustering techniques in terms of information conservation.

The MAgPIE model is used for these comparisons. First, we have generalized the model structure to be able to simulate in various spatial aggregations instead of being restricted to a single resolution of 3.0° as it was in previous versions of the model. Second, we have implemented spatial aggregation methods (grid-based and clustering-based) to merge input data to these aggregations (together this allows the model to be run at various spatial aggregations). Third, we have implemented an interpolation methodology to downscale clustered outputs back to the grain size of the input data. Last, we have used this implementation to compare the standard aggregation method using a regular grid with hierarchical and non-hierarchical clustering methods.

2. Methods

2.1. Model implementation

MAgPIE is a recursive cost-minimizing equilibrium model with three involved scales: a global scale representing global markets, a regional scale of 10 world regions[1] representing specific economic development, demands and technology levels, and a local scale representing farming decisions based on spatially varying production parameters, as for instance potential yields and water availability (see the mathematical model description in the supplementary online material for more details). The model is written in GAMS [5] extended with scripts for file manipulations written in PHP [3] and scripts for calculations written in R [22] and Python [24]. Since GAMS does not allow for calculating sets and therefore cannot handle inputs with varying grain sizes, a PHP script is executed before GAMS is started. The PHP script organizes the aggregation of the original input data set and rewrites the sets in the GAMS source code according to the chosen grain size. The aggregation of input data itself is done in R, either by using a regular grid or clustering aggregation. After execution of the GAMS model, the clustered data is downscaled to the grain size of the original input data using another R and Python script.

The unprocessed input data has a grain size of 0.5° (i.e. 30 arcminutes of longitude and latitude). Each cell contains information on the potential yields of 20 different crops (rainfed and irrigated)[2], crop-specific demands for irrigation water, the total amount of water available for irrigation (all calculated by the "Lund-Potsdam-Jena global vegetation model with managed land" (LPJmL) [4]), total cropland area and total land available for additional cropland expansion [16].

2.2. Aggregation methods

For aggregation two approaches are implemented: (A) an aggregation based on regular grids and (B) an aggregation using clustering methods. In any case only cells that belong to the same world region are aggregated together.

In the case of regular grids a grain size is chosen (coarser than the original grain size of 0.5°) and input data cells lying in the same coarser cell are either summed up or (weighted) averaged depending on the type of data. Yields are averaged using the total crop share of a cell as weight; the amount of available water per cell is summed up; the required amount of water for each crop is also crop-area weighted averaged; and crop shares are cell-area weighted averaged.

For the clustering methods the target grid is chosen depending on the data to be aggregated. All clustering methods have in common that clusters are built on some kind of multivariate distance measure between data elements. Every cell is represented by its data and each data set is standardized over all grid cells to get a balanced weighting across data sets. The distance between cells is based on the similarity of data, for instance cells with similar yields are close to each other, whereas large differences in yields lead to high distances between cells (not to be mistaken for physical distance). The distance is measured in the n-dimensional space spanned by the n data sets which define the characteristics of each cell. Because of regional separation, every cluster belongs to exactly one region. In contrast to grid-based aggregation, clusters are not necessarily connected to a single, contiguous spatial location. It can happen that one cluster is divided into two or more disjoint patches distributed over the region. Furthermore, clustering does not increase the grain size, since the smallest unit which a clustered data set can contain is still one cell of the original data set. Instead of increasing the grain size, cluster methods try to reduce the number of spatial units by combining cells with similar characteristics.

Figure 1 and Figure 2 illustrate the different aggregation approaches schematically. Using a regular grid the procedure is quite simple (Figure 1). In the example shown the initial data set has 81 cells with a grain size of 1x1 and an extent of 9x9. The values of each cell are indicated with colors. Increasing the grain size

[1] AFR = Sub-Sahara Africa, CPA = Centrally Planned Asia (incl. China), EUR = Europe (incl. Turkey), FSU = Former Soviet Union, LAM = Latin America, MEA = Middle East and North Africa, NAM = North America, PAO = Pacific OECD (Australia, Japan and New Zealand), PAS = Pacific Asia, SAS = South Asia (incl. India)

[2] wheat, rice, maize, millet, pulses, cotton, potato, sugar beet, sugar cane, cassava, sunflower, soybean, groundnut, palm oil, rapeseed, bioenergy grasses, bioenergy trees, fodder, pasture, others

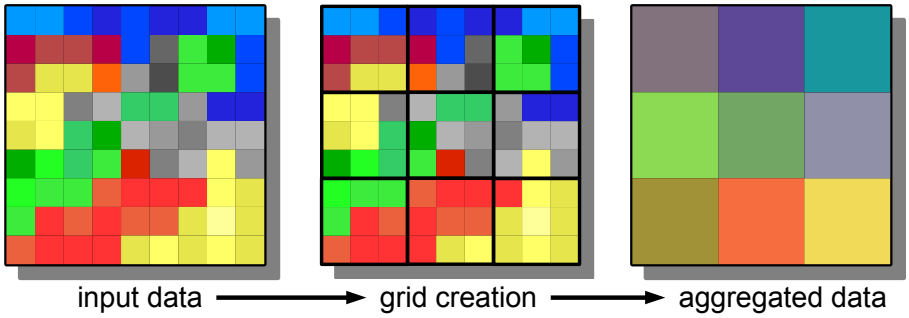

Figure 1: Regular grid aggregation

to 3x3 creates the regular grid, which is used for aggregation. The colors of the new data set are derived by averaging the colors of all elements within a segment. Depending on the homogeneity of a segment the resulting color is either similar to the colors of the input data set (lower-right segment) or quite different (upper-left segment). In the given example the initial color distribution is hardly visible in the aggregated diagram, which indicates a significant loss of effective resolution. Schmit et al. [27] give good examples of the biases and problems this kind of information loss can lead to.

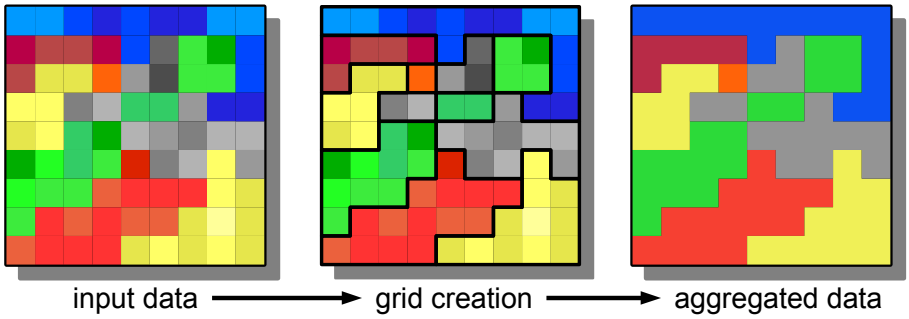

Figure 2: Cluster-based aggregation

Figure 2 schematically shows cluster-based aggregation starting with the same input data set as in Figure 1. Through clustering, the input data itself determines the merging of cells. In the grid creation step, cells are merged on the basis of their similarity. In contrast to regular grid aggregation the cluster size is variable. This allows clusters to be formed containing only one cell (e.g. the orange cell in the schematic diagram). The grain size of the data set is not increased. Merged cells do not have to be neighbors. One cluster may be distributed over the whole data set (e.g. the green or the yellow cluster). However, this makes it impossible to reasonably link a cluster with a single, spatial coordinate. Therefore, data calculated on that aggregation level (e.g. model outputs) have to be downscaled again before they can be used for spatially explicit calculations. The deviation between the input and the aggregated data set shows that some data is lost. However, the characteristics of the initial data set are still visible, indicating a relatively high level of effective resolution conservation. In this context it is also interesting to note that the cluster example shown here only used 7 clusters, while the regular grid aggregation was performed with 9 segmentation units.

Besides regular grid aggregation, three different clustering methods are applied in the following: k-means clustering, top-down hierarchical clustering and bottom-up hierarchical clustering.

2.3. Clustering methods

2.3.1. k-means clustering

k-means is a method to partition n cells with data $x_1, ..., x_n$ into k clusters $\mathbf{S} = \{\mathbf{S}_1, ..., \mathbf{S}_k\}$ with mean values $\mu_1, ..., \mu_k$ by minimizing the within-cluster sum of squares (WCSS) (Equation 1) [12, 10].

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{x_j \in \mathbf{S}_i} \|x_j - \mu_i\|^2 \tag{1}$$

The problem can be solved analytically, but because of its high computational intensity typically heuristic algorithms are used. We use the heuristic implementation offered by the pycluster python module [14]. This approach is relatively fast and delivers accurate results but is not fully deterministic, which limits its reproducibility. Another important characteristic of the k-means algorithm is its weight sensitivity [1]: the cluster pattern is dependent on the weighting factors of each cell and changes if the weighting factors are changed. In the given case this causes an unwanted distortion as the used algorithm only allows for equally weighted cells, which does not properly reflect the unequally sized input cells. Applied to a global model with distinct world regions such as MAgPIE one has to face the additional constraint that all cells within one cluster have to belong to the same world region. For k-means clustering this is assured by initializing separate clustering processes for each region. The number of clusters for each region is chosen in proportion to the size of the region.

2.3.2. Hierarchical clustering

Hierarchical clustering methods can be classified in two types: top-down and bottom-up. Both create a hierarchy based on the distances between data elements. Top-down methods start with one cluster which is disaggregated step by step. Bottom-up methods start with cells which are aggregated stepwise. Here the distances between clusters are measured with a euclidean metric and the complete-linkage method, which measures the distance between clusters as the maximum distance between two cells of both clusters. Complete linkage is weight robust and therefore free from the distortions described for the k-means method [1]. Single linkage (minimum distance between cells of two clusters), average linkage (average distance between cells of both clusters) and centroid linkage (distance between centroids of two clusters) were also tested but excluded from this analysis, since they produced at best equivalent but in most cases significantly inferior results in terms of our evaluation measures (Section 2.5). For bottom-up hierarchical clustering the implementation of the pycluster module [14] is used. For top-down clustering we developed an own implementation in which a cluster is split by first separating cells with the longest distances in between. All remaining cells are then distributed accordingly.

In contrast to k-means clustering, hierarchical clustering is fully deterministic and results produced with different numbers of clusters can be merged easily based on its hierarchy. However, the hierarchy also limits the adaptivity of the clusters to the data, because of lock-in effects: Clustering decisions made earlier in the hierarchy cannot be reversed. Separate clustering processes for each region were started and merged to one hierarchy tree based on distance information (clusters with smallest distances are combined first). This approach allows for distributing clusters in a globally optimal way, while preventing cross-regional clustering.

2.4. Downscaling

Two significant problems arise when dealing with clustered data. First, the clusters do not necessarily have a spatial meaning, since clusters can be distributed over the whole map. Second, only outputs directly derived from input data involved in the clustering procedure are meaningful. With respect to values that were not part of the clustering, model outputs in each cluster are only the mean of a potentially strongly heterogeneously distributed value and a significant amount of information is lost. Both problems can be solved by downscaling. The problem of missing spatial information is solved by inversion of the aggregation process, whereas the information loss in the second case is compensated by additional knowledge about the distribution used within the downscaling process.

The downscaling can be divided into two steps. In step one, data are processed using trivial downscaling rules. Trivial downscaling is either downscaling by giving each cell the value of its associated cluster or downscaling by partitioning a summed-up value proportionately to data that are known already in the higher resolution. In all other cases downscaling rules are non-trivial. Values which require non-trivial downscaling are not processed directly. Instead, such a value is split into components with trivial downscaling rules, which are then downscaled and recombined to the downscaled counterpart of the initial value. One example is cellular, crop-specific data on cropland shares. First, these shares are split into a term containing crop-specific total areas and a term containing the total cluster areas. Second, crop-specific total areas are downscaled by partitioning the data proportionally to total cell area (total cell area is constant and known at 0.5° grain size and, therefore, does not need to be downscaled). Third, the downscaled, crop-specific areas are divided by the total area to reconstruct the original shares at 0.5° grain size.

2.5. Evaluation

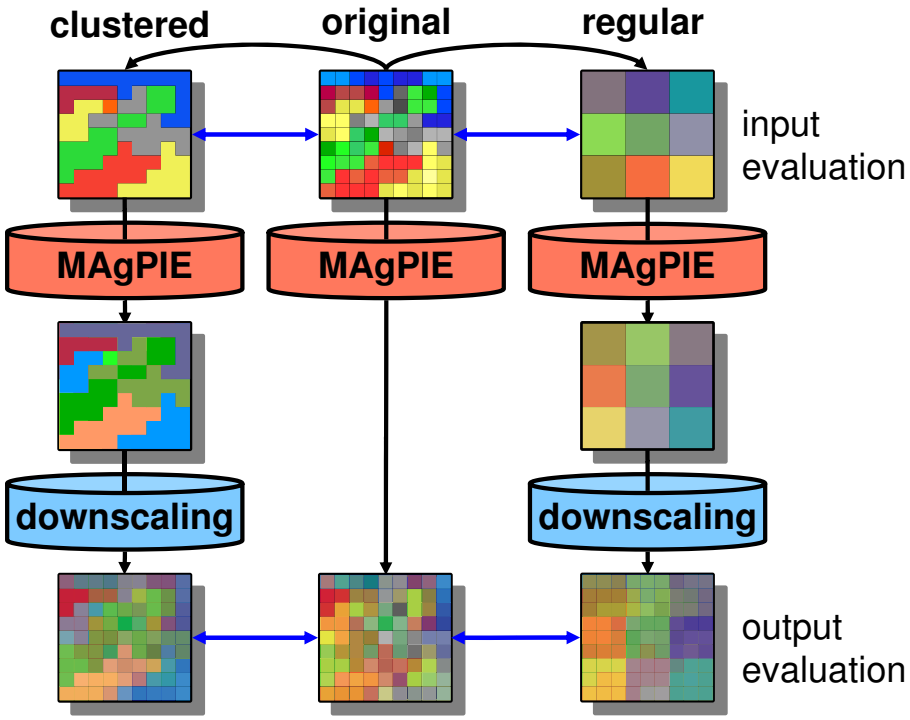

Figure 3: experiment setup

To test the performance of the different aggregation methods we performed an experiment with an input and an output evaluation phase (Figure 3): In the first phase we aggregate a reference data set to a given grain size (using a regular grid method or clustering methods) and disaggregate it again by giving each cell in the reference resolution the value of the corresponding cluster to which it was assigned. Subsequently the loss of information is measured by comparing this newly derived data set (Figure 3, upper corners) with the reference data set (Figure 3, upper mid place). This input evaluation tells us how much information is lost due to the aggregation itself. In the second phase we use the same aggregation procedures, but this time we use the reference data and the aggregated data as input for a model run with the MAgPIE model. Next the outputs of the run with reference data and the run with aggregated data (based on a regular grid or clustered data) are compared and the loss of information due to aggregation is measured by assessing the model outputs. Non-cellular model outputs (global or regional) are compared directly, cellular outputs are downscaled beforehand to the grain size of the reference data.

For standard MAgPIE runs all cellular inputs are derived from 0.5° source data. However, for the performance tests we also need the ability to run simulations with the reference data. This was not possible with 0.5° data as it exceeded available computational resources. To circumvent this issue we created a new reference data set with a grain size of 2.0° and 4669 cells, which can be used directly for simulations with MAgPIE.

The choice of aggregation levels for testing is driven by two conditions. First, each aggregation should have a regular grid counterpart, which means that 180°, the maximum extent in north-south direction, must be a multiple of the chosen grain size. Second, the numbers of data elements should be equally spaced on a logarithmic scale. This led to 17 aggregation levels with 14 to 3772 data elements (180° to 2.25° grid size) using regular grids, k-means clustering, hierarchical bottom-up clustering and hierarchical top-down clustering.

To measure the quality of the different aggregations $i = 1..n$, two measures $d_1(i)$ and $d_2(i)$ are applied. Both measures can be written in general as the mean of normalized distances (Equation 2). The measures are very unlike in their calculation which was intended to give a broader view of the topic by allowing comparisons of results among measures.

$$d_k(i) = \frac{1}{m} \sum_{s=1}^{m} \frac{\hat{d}_k(X_{i,s}, X_{0,s})}{\max_{j=1}^{n} \hat{d}_k(X_{j,s}, X_{0,s})} \tag{2}$$

The quality of the i-th aggregation is calculated by taking the mean of the distances $\hat{d}_k$ between aggregated data X_j and reference data X_0 for all data sets $s = 1..m$ used for comparison (e.g. rainfed yield of maize is one data set, available discharge of each cell is another one), normalized by the maximum distance observed for each data set. This approach delivers values between 0 and 1, with 0 in the case of a perfect match between reference and aggregated data and 1 in the case of the maximum observed deviation between the data sets. The measures can be interpreted as measures for the effective resolution of a data set, where values close to 0 indicate an effective resolution similar to the original data set and values close to 1 indicate high losses in effective resolution. For the calculation two distance measures $\hat{d}_1$ and $\hat{d}_2$ were applied:

1. The euclidean distance $\hat{d}_1$ between two data sets Y and Z with l data points:

$$\hat{d}_1(Y, Z) = \|Y - Z\| = \sqrt{\sum_{k=1}^{l} (y_k - z_k)^2} \tag{3}$$

   This measures the similarity between original and aggregated data set: the more information is conserved, the lower the final value will be.

2. The mutual information distance d_2. For calculation of mutual information distances we use the R package "bioDist"[8]. Mutual information m is a nonlinear measure for the mutual dependence of two variables Y and Z (Equation 4).

$$m(Y, Z) = \sum_{y \in Y} \sum_{z \in Z} p(y, z) \log \left(\frac{p(y, z)}{p(y)p(z)} \right) \tag{4}$$

   $p(y, z)$ is the joint probability density function of Y and Z, $p(y)$ and $p(z)$ are the marginal probability density functions of Y and Z respectively. Mutual information is a measure for the amount of information which is shared by both variables. So it should be high if the aggregation conserves much information of the original dataset and should be low otherwise. The mutual information distance $\hat{d}_2$ is calculated by applying the transformation of equation 5 to the mutual information m, which was proposed by Joe [15].

$$\hat{d}_2(Y, Z) = 1 - \sqrt{1 - \exp(-2m(Y, Z))} \tag{5}$$

   Like $\hat{d}_1$, $\hat{d}_2$ also measures the similarity between original and aggregated data set.

description	n	unit
crop-specific water demand for irrigation	20	mm/yr
water available for irrigation	1	$10^6 m^3/yr$
crop-specific rainfed yields	20	$t/(ha \cdot yr)$
crop-specific yields under optimal irrigation	20	$t/(ha \cdot yr)$

Table 1: List of cellular model inputs with corresponding number of values per cell (n)

description	level	n	unit
total costs of production	global	1	$10^6 US\$/yr$
total production value	global	1	$10^6 US\$/yr$
gross area of converted land	global	1	$10^6 ha/yr$
crop-specific area	global	20	$10^6 ha$
average of technological change rates	global	1	$1/yr$
crop-specific area	regional	20	$10^6 ha$
average technological change rates	regional	1	$1/yr$
supply-demand balance	regional	1	–

Table 2: List of non-cellular model outputs with its spatial level and number of values per spatial unit (n)

description	n	unit
total cropland share of total area	1	–
pasture land share of total area	1	–
available land share of total area	1	–
crop-specific land use shares of total area	20	–
shadow prices of land	1	$US\$/ha$
producer rent including land rent	1	$10^6 US\$/yr$

Table 3: List of cellular model outputs with corresponding number of values per cell (n)

Measures are applied separately on cellular model inputs, which are also involved in the clustering procedure (Table 1), and a selection of default non-cellular (global or regional) and cellular model outputs (Table 2, Table 3). Since the mutual information measure d_2 only works well for large data sets, it is only applied on cellular data, whereas d_1 is also applied on non-cellular data.

To test the significance of measured differences in quality between aggregation types the Wilcoxon signed-rank test is used [36]. Comparable to the paired Student's t-test, the Wilcoxon test is used to check whether two related samples are part of the same distribution or not. We use it instead of the t-test as it does not require the assumption of normally distributed data.

3. Results

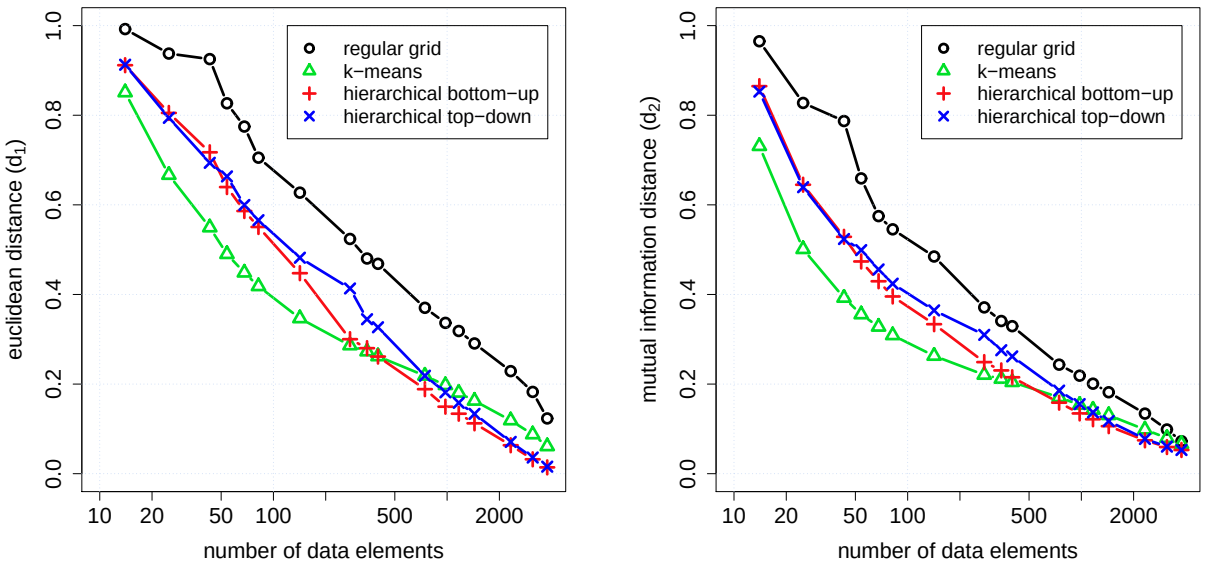

Figure 4: Input evaluation with normalized euclidean distance d_1 (left) and normalized mutual information distance d_2 (right)

Applying the measures d_1 and d_2 and a Wilcoxon one-sided, signed-rank test, the clustering methods show in the input evaluation significantly better results over the full tested range compared to the standard aggregation using regular grids (Figure 4, Table 4). However, when we compare the different clustering methods the picture is more diverse and the results depend on level of aggregation. For further analysis, three overlapping classes were extracted from the full tested range of 14 - 3772 data elements: coarse resolutions ranging from 14 to 346 data elements, medium resolutions ranging from 54 to 1167 data elements and fine resolutions ranging from 346 to 3772 data elements.

Overall, k-means shows the best results in the input evaluation for coarse resolutions, for which it behaves significantly better than both hierarchical methods. For medium resolutions k-means still delivers significantly better results compared to top-down hierarchical clustering, but non-significant advantages compared to bottom-up clustering. In the case of fine resolutions bottom-up clustering shows the best performance and is significantly better than top-down clustering in both quality measures and significantly better than k-means clustering in measure d_1 (Table 4). Seen over the full range of aggregated resolutions we considered top-down clustering behaves slightly worse than bottom-up clustering, with nearly identical behavior for coarse resolutions but significantly worse results for medium and fine resolutions.

In the output evaluation the picture changes (Figure 5, Table 5). All three clustering methods perform significantly better than aggregation by regular grids over the full range of resolutions, but magnitudes of differences and significance levels are lower. Whereas the quality ratings of input data varied over nearly

H_0		full range		coarse resolutions		medium resolutions		fine resolutions	
h ≥ r	d1	0.000	***	0.004	**	0.001	***	0.001	***
	d2	0.000	***	0.004	**	0.001	***	0.001	***
k ≥ r	d1	0.000	***	0.004	**	0.001	***	0.001	***
	d2	0.000	***	0.004	**	0.001	***	0.001	***
t ≥ r	d1	0.000	***	0.004	**	0.001	***	0.001	***
	d2	0.000	***	0.004	**	0.001	***	0.001	***
r ≥ h	d1	1.000		1.000		1.000		1.000	
	d2	1.000		1.000		1.000		1.000	
k ≥ h	d1	0.112		0.004	**	0.138		0.993	
	d2	0.044	*	0.004	**	0.053		0.903	
t ≥ h	d1	0.999		0.945		1.000		1.000	
	d2	0.999		0.961		1.000		0.999	
r ≥ k	d1	1.000		1.000		1.000		1.000	
	d2	1.000		1.000		1.000		1.000	
h ≥ k	d1	0.897		1.000		0.884		0.010	**
	d2	0.960		1.000		0.958		0.116	
t ≥ k	d1	0.993		1.000		0.993		0.539	
	d2	0.997		1.000		0.998		0.722	
r ≥ t	d1	1.000		1.000		1.000		1.000	
	d2	1.000		1.000		1.000		1.000	
h ≥ t	d1	0.001	***	0.074		0.001	***	0.001	***
	d2	0.002	**	0.055		0.001	***	0.002	**
k ≥ t	d1	0.009	**	0.004	**	0.010	**	0.500	
	d2	0.004	**	0.004	**	0.003	**	0.312	

Table 4: Input evaluation of hierarchical bottom-up (h), hierarchical top-down (t), k-means (k) and regular grid (r) aggregation for the full range of resolutions (14-3772 data elements), coarse resolutions (14-346 data elements), medium resolutions (54-1167 data elements) and fine resolutions (346-3772 data elements) using p-values of a Wilcoxon signed-rank test applied to the normalized euclidean distance d_1 and normalized mutual information distance d_2 (significance levels * p≥95%, ** p≥99%, *** p≥99.9%).

H_0		full range		coarse resolutions		medium resolutions		fine resolutions	
h ≥ r	d1	0.000	***	0.008	**	0.002	**	0.001	***
	d2	0.000	***	0.012	*	0.003	**	0.001	***
k ≥ r	d1	0.003	**	0.039	*	0.002	**	0.014	*
	d2	0.000	***	0.004	**	0.001	***	0.001	***
t ≥ r	d1	0.000	***	0.039	*	0.024	*	0.003	**
	d2	0.000	***	0.020	*	0.002	**	0.001	***
r ≥ h	d1	1.000		0.996		0.999		1.000	
	d2	1.000		0.992		0.998		1.000	
k ≥ h	d1	0.998		0.680		0.884		1.000	
	d2	0.920		0.008	**	0.688		0.997	
t ≥ h	d1	1.000		0.926		0.997		1.000	
	d2	0.998		0.680		0.993		1.000	
r ≥ k	d1	0.997		0.973		0.999		0.990	
	d2	1.000		1.000		1.000		1.000	
h ≥ k	d1	0.002	**	0.371		0.138		0.001	***
	d2	0.087		0.996		0.348		0.005	**
t ≥ k	d1	0.274		0.629		0.990		0.188	
	d2	0.500		1.000		0.968		0.116	
r ≥ t	d1	1.000		0.973		0.981		0.998	
	d2	1.000		0.988		0.999		1.000	
h ≥ t	d1	0.001	***	0.098		0.005	**	0.001	***
	d2	0.003	**	0.371		0.010	**	0.001	***
k ≥ t	d1	0.741		0.422		0.014	*	0.839	
	d2	0.518		0.004	**	0.042	*	0.903	

Table 5: Output evaluation of hierarchical bottom-up (h), hierarchical top-down (t), k-means (k) and regular grid (r) aggregation for the full range of resolutions (14-3772 data elements), coarse resolutions (14-346 data elements), medium resolutions (54-1167 data elements) and fine resolutions (346-3772 data elements) using p-values of a Wilcoxon signed-rank test applied to the normalized euclidean distance d_1 and normalized mutual information distance d_2 (significance levels * p≥95%, ** p≥99%, *** p≥99.9%).

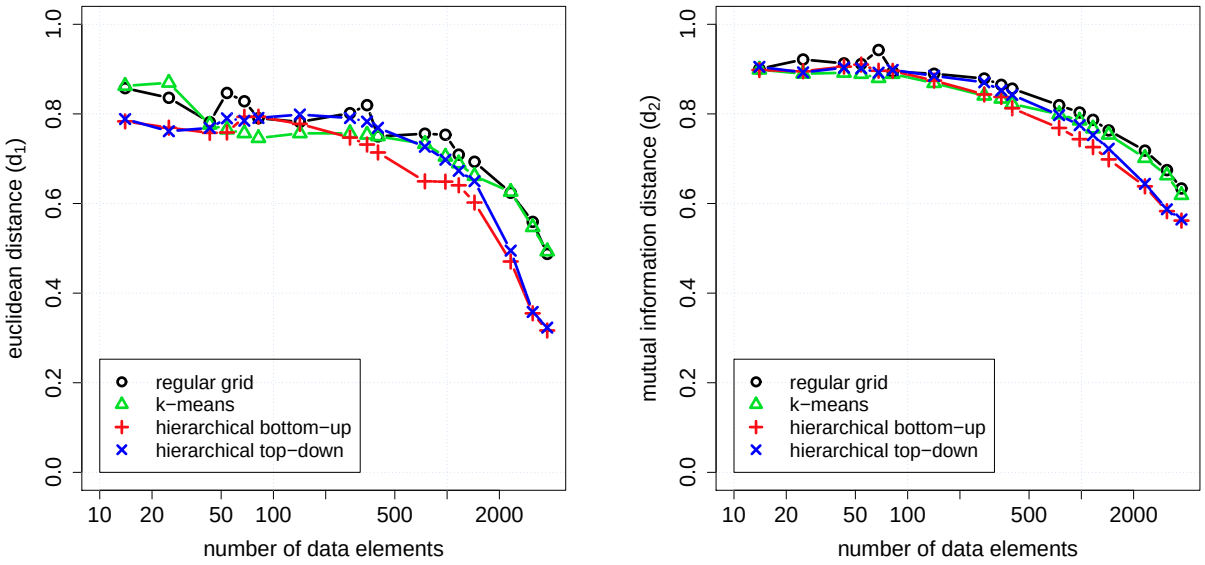

Figure 5: Output evaluation with normalized euclidean distance d_1 (left) and normalized mutual information distance d_2 (right)

the whole range from 0 to 1, the ratings of output data only range from 0.3 for d_1 or 0.5 for d_2 to 0.9. This reduced variance in quality measures reflects an increased variance in rankings of the different aggregation experiments across different output data sets. That means the distance rankings of different aggregation methods and number of clusters depend more on the chosen set of output data, whereas the rankings in input data primarily depend on the number of clusters and aggregation method. Hence, some outputs might get better with k-means clustering, other outputs with hierarchical clustering. It even happens that a less aggressive aggregation increases the distance between the reference and the aggregated case (sections with positive slopes in figure 5).

Comparing the results of input and output evaluation, we see that k-means shows noticeable changes in overall performance: whereas in the input evaluation k-means shows a significantly better performance than hierarchical clustering for coarse resolutions, this superiority mostly vanishes in the output evaluation. Whereas d_2 still reports significantly better results for k-means at coarse resolutions, this effect cannot be observed for d_1 anymore. In fact the k-means output evaluation shows in d_1 for coarse resolutions an even worse behavior than regular grids. Comparing the hierarchical clustering methods, the results of input evaluation are similar to those of the output evaluation: the two methods produce similar results with slight, but significant advantages for the bottom-up approach (significant for medium and fine resolutions, non-significant for coarse resolutions).

Analyzing non-cellular and cellular outputs separately one finds that in general, results for non-cellular data are much fuzzier than those for cellular data (Figure 6, Table 6). For hierarchical clustering methods both output evaluations still show significant quality improvements compared to regular grids. However, for k-means the results only remain significant for cellular data. For non-cellular data regular grids deliver even better results than k-means in the case of fine resolutions.

4. Discussion

Agricultural land-use models combine processes across different scales. While some economic processes, like commodity trade, occur at the global scale, ecological parameters and farming decisions come into play at regional to local scales. Linking these scales is in land-use modeling - as in many other research areas - one of the important issues that have to be faced [35, 6, 11]. In this paper we addressed the problem of

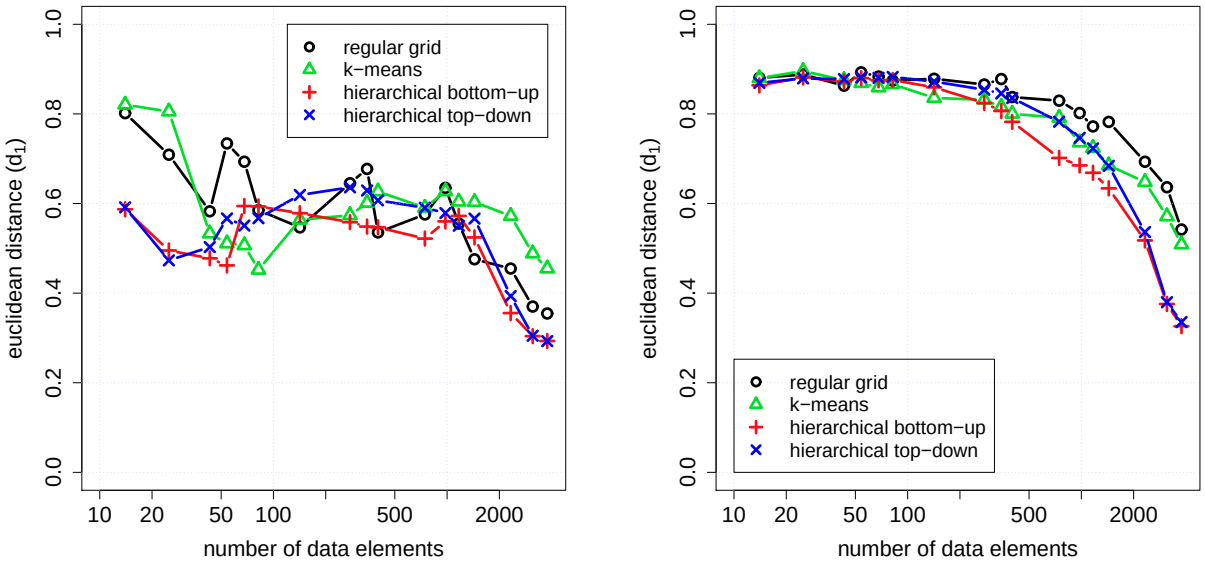

Figure 6: Output evaluation with normalized euclidean distance d_1 for non-cellular outputs (left) and cellular outputs d_2 (right)

information loss due to aggregation which often occurs in this context. Aggregation through clustering is one option to tackle this issue and to enhance model precision.

In our experiment setup all presented clustering methods deliver at least equal but mostly better results in terms of information conservation compared to standard aggregation based on regular grids. While the application of clustering methods is superior to regular grid aggregation in all the situations tested, the explicit choice of the clustering method depends on the application and number of clusters. In our investigations, which compare original and aggregated data sets directly, two cases can be distinguished. Working with a few clusters (less than 500 clusters in the case of 2.0° grain size), k-means clustering delivered the lowest information losses due to aggregation and the highest effective resolution conservation. Moving up to a higher number of clusters, hierarchical bottom-up clustering with complete linkage became the better choice. Hierarchical top-down clustering behaved slightly worse over the whole range than hierarchical bottom-up. It seems that the advantage of top-down clustering in not having a significant lock-in effect for a few clusters is overshadowed by other effects. Since the complete linkage method only determines that the two cells with the highest distance within a cluster have to be separated in an iteration step, but does not make any decisions on the mapping of the remaining cells, several representations of complete-linkage top-down clustering are possible. So the bad performance of top-down clustering may just be an effect of the chosen algorithm and not of the method itself. A counterexample can be found for instance in Steinbach et al. [28] which describes clustering with a mixture of hierarchical top-down and k-means algorithm ("bisecting k-means").

In the output evaluation the ranking between different clustering methods changes. The advantages of k-means clustering for a few data elements compared to hierarchical bottom-up are no longer visible and hierarchical bottom-up clustering now shows the best results over the whole range. A possible explanation for this behavior is that k-means clustering is based on euclidean distances, whereas both hierarchical clustering methods use a complete linkage approach which is a representation of the infinity norm. This leads to significant differences in the handling of outliers. In the case of the infinity norm outliers are preferably used as single clusters since only the maximum distance between two clusters is accounted for and not the sum of all connections between them. In the case of a euclidean distance, outliers will more often be embedded in an existing cluster since this method not only accounts for the maximum distance but for all distances between cells of both clusters (see Appendix A for more details). Looking at the aggregated

H_0		full range		coarse resolutions		medium resolutions		fine resolutions	
h ≥ r	non-cellular	0.001	**	0.020	*	0.042	*	0.014	*
	cellular	0.000	***	0.039	*	0.002	**	0.001	***
k ≥ r	non-cellular	0.627		0.098		0.138		0.968	
	cellular	0.000	***	0.055		0.001	***	0.001	***
t ≥ r	non-cellular	0.032	*	0.020	*	0.188		0.312	
	cellular	0.001	***	0.191		0.007	**	0.001	***
r ≥ h	non-cellular	0.999		0.988		0.968		0.990	
	cellular	1.000		0.973		0.999		1.000	
k ≥ h	non-cellular	0.995		0.809		0.784		1.000	
	cellular	0.994		0.371		0.812		1.000	
t ≥ h	non-cellular	0.985		0.809		0.968		0.993	
	cellular	1.000		0.992		1.000		1.000	
r ≥ k	non-cellular	0.391		0.926		0.884		0.042	*
	cellular	1.000		0.961		1.000		1.000	
h ≥ k	non-cellular	0.005	**	0.230		0.246		0.001	***
	cellular	0.007	**	0.680		0.216		0.001	***
t ≥ k	non-cellular	0.122		0.629		0.920		0.042	*
	cellular	0.627		0.945		0.997		0.312	
r ≥ t	non-cellular	0.972		0.988		0.839		0.722	
	cellular	0.999		0.844		0.995		1.000	
h ≥ t	non-cellular	0.017	*	0.230		0.042	*	0.010	**
	cellular	0.000	***	0.012	*	0.001	***	0.001	***
k ≥ t	non-cellular	0.888		0.422		0.097		0.968	
	cellular	0.391		0.074		0.005	**	0.722	

Table 6: Output evaluation separated for cellular and non-cellular outputs of hierarchical bottom-up (h), hierarchical top-down (t), k-means (k) and regular grid (r) aggregation for the full range of resolutions (14-3772 data elements), coarse resolutions (14-346 data elements), medium resolutions (54-1167 data elements) and fine resolutions (346-3772 data elements) using p-values of a Wilcoxon signed-rank test applied to the normalized euclidean distance d_1 (significance levels * p≥95%, ** p≥99%, *** p≥99.9%).

data itself it is preferable to use a well-balanced distance measure, such as the euclidean distance, since it accounts for any distance of any cell. Looking specifically at outputs of optimization models, extreme values as supplied by outliers are relevant for the optimization result. Hence, in order to reconstruct the result produced with the reference data itself it is preferable to describe outliers as single clusters. This will reduce the loss of information which is most relevant for the modeling output. Consequently, the complete linkage method is the best choice for preserving information most relevant for the optimization. When working with real data,the correctness of outliers in particular is often extremely doubtful. From that point of view k-means clustering while reducing the similarity between outputs of reference data and aggregated data, is probably reducing the biases caused by flawed outliers. Hence, the choice of clustering method is also a question of the reliability of the input data.

The role of outliers can also explain the fact that our results for k-means clustering are even worse for non-cellular outputs than for cellular outputs. Whereas in the case of cellular outputs each single cell still plays an important role for the total result, non-cellular outputs can be much more influenced by outliers-induced effects. But also for all other clustering methods, results are less significant for non-cellular data than for cellular ones. This is also due to the fact that for non-cellular data, errors from aggregation more frequently cancel each other out. This could explain the high randomness in the quality of results and also why even the reduction of clusters can improve results.

5. Conclusion

In land use modeling, the combination of high-resolution data with high-complexity models remains a challenge. One approach to deal with it is clustering. Our tests show that especially hierarchical bottom-up clustering can lead to a significant reduction of information loss due to aggregation. Other clustering methods (e.g. hierarchical top-down or k-means clustering) also increase the quality of aggregated data in our experiments, but to a lesser extent.

All the results shown have to be treated with care as they were derived with a very specific experiment setup so that generic validity for most outcomes is not given. Especially the comparison of different cluster techniques is highly dependent on the specific case and the picture would probably be different for a different experiment setup. However, our study produced some robust outcomes. Our results show that there is a significant difference between the results of the input and output evaluation: aggregations which show a high degree of information conservation in inputs can nevertheless cause huge differences in model outputs compared to runs with the original data. This indicates that the information used for the inputs is not all equally relevant for the model calculations. In our specific case it seems that especially the conservation of extreme values is of major importance for producing results resembling those achieved with the original data. However, what kind of data is most relevant for the model outputs again strongly depends on the kind of model which is applied. For the MAgPIE model our finding is that hierarchical bottom-up clustering is the most favorable method as it performed best in the output evaluation. Nevertheless, k-means clustering is also useful in cases when it is important to have more equally sized clusters or when outliers in the cellular input data are expected to be misleading. Therefore, both types of aggregation are currently in use in MAgPIE.

Clustering does not solve the scale problem in land-use models, but - independent of the particular choice of clustering method - it helps increase model accuracy and reduce information losses within the aggregation process. Compared to spatial aggregation with regular grids, in our analysis clustering always showed results closer to the results achieved with inputs and simulation on the smallest possible grain size. Thus it is the better choice for aggregation.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreements n°265170 (ERMITAGE) and n°265104 (VOLANTE). We are grateful for many valuable discussions with our colleagues Katharina Waha, Christoph Müller and Susanne Rolinski. Furthermore, we would like to thank Alison Schlums for proof-reading this paper. We wish to thank our reviewers for their valuable comments which substantially improved the paper.

[1] M. Ackerman, S. Ben-David, S. Branzei, and D. Loker. Weighted clustering. In *Proc. 26th AAAI Conference on Artificial Intelligence*, 2012. URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/download/4909/5492`.

[2] D. M Adams, R. J Alig, J. M. Callaway, B. A McCarl, and S. M Winnett. The forest and agricultural sector optimization model (FASOM): model structure and policy applications. *Notes*, 1996.

[3] S. S Bakken, A. Aulbach, E. Schmid, J. Winstead, L. T Wilson, R. Lerdorf, and Z. Suraski. PHP manual. *PHP Documentation Group*, pages 12–04, 2004.

[4] A. Bondeau, P.C. Smith, S. O. N. Zaehle, S. Schaphoff, W. Lucht, W. Cramer, D. Gerten, H. Lotze-Campen, C. Müller, and M. Reichstein. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology*, 13(3):679–706, 2007.

[5] A. Brook, D. Kendrick, and A. Meeraus. GAMS, a user's guide. *ACM SIGNUM Newsletter*, 23(3-4):11, 1988.

[6] D. W Cash and S. C Moser. Linking global and local scales: designing dynamic assessment and management processes. *Global Environmental Change*, 10(2):109–120, 2000. ISSN 0959-3780.

[7] T. Dirnböck, P. Bezák, S. Dullinger, H. Haberl, H. Lotze-Campen, M. Mirtl, J. Peterseil, S. Redpath, S. J. Singh, J. Travis, et al. Scaling issues in long-term socio ecological biodiversity research. a review of european cases. Technical report, Social Ecology Working Paper, 2008.

[8] R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim. Distance measures in DNA microarray data analysis. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 189–208, 2005.

[9] C. C Gibson, E. Ostrom, and T. K Ahn. The concept of scale and the human dimensions of global change: a survey. *Ecological Economics*, 32(2):217–239, 2000. ISSN 0921-8009.

[10] J. A Hartigan. *Clustering algorithms*. Wiley New York, 1975.

[11] L. D.D Harvey. Upscaling in global change research. *Climatic Change*, 44(3):225–263, 2000. ISSN 0165-0009.

[12] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
[13] M. Heistermann, C. Müller, and K. Ronneberger. Land in sight? achievements, deficits and potentials of continental to global scale land-use modeling. *Agriculture, Ecosystems & Environment*, 114(2-4):141–158, 2006.
[14] M. J. L. De Hoon, S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. *Bioinformatics*, page 781, 2004.
[15] H. Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405): 157–164, 1989.
[16] M. Krause, H. Lotze-Campen, and A. Popp. Spatially-explicit scenarios on global cropland expansion and available forest land in an integrated modelling framework. *27th International Conference of Agricultural Economists (IAAE), Bejing*, 2009.
[17] J. Krüger. Hierarchie und skalen als ordnungsprinzipien für die abbildung des objektumfanges und der beschreibung kohärenter strukturen. *Leipziger Geowissenschaften*, 18:123–138, 2007.
[18] A. Letourneau, P. H. Verburg, and E. Stehfest. A land-use systems approach to represent land-use dynamics at continental and global scales. *Environmental Modelling & Software*, 33(0):61–79, July 2012. ISSN 1364-8152. doi: 10.1016/j.envsoft.2012.01.007. URL `http://www.sciencedirect.com/science/article/pii/S136481521200014X`.
[19] H. Lotze-Campen, C. Müller, A. Bondeau, S. Rost, A. Popp, and W. Lucht. Global food demand, productivity growth and the scarcity of land and water resources: a spatially explicit mathematical programming approach. *Agricultural Economics*, 39(3):325–338, 2008.
[20] H. Lotze-Campen, A. Popp, T. Beringer, C. Müller, A. Bondeau, S. Rost, and W. Lucht. Scenarios of global bioenergy production: The trade-offs between agricultural expansion, intensification and trade. *Ecological Modelling*, page (online first), 2009.
[21] A. Popp, H. Lotze-Campen, and B. Bodirsky. Food consumption, diet shifts and associated non-CO2 greenhouse gases from agricultural production. *Global Environmental Change*, 2010.
[22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.
[23] M. W Rosegrant, S. Meijer, and S. A Cline. International model for policy analysis of agricultural commodities and trade (IMPACT): model description. *IFPRI, Washington, DC*, 2002.
[24] G. Van Rossum and F. L Drake Jr. *Python reference manual*. iUniverse, 2000.
[25] R. D Sands and M. Leimbach. Modeling agriculture and land use in an integrated assessment framework. *Climatic Change*, 56(1):185–210, 2003.
[26] R. Schaldach, J. Alcamo, J. Koch, C. Kölking, D. M. Lapola, J. SchÃ$\frac{1}{4}$ngel, and J. A. Priess. An integrated approach to modelling land-use change on continental and global scales. *Environmental Modelling & Software*, 26(8):1041 – 1051, 2011. ISSN 1364-8152. doi: 10.1016/j.envsoft.2011.02.013. URL `http://www.sciencedirect.com/science/article/pii/S1364815211000570`.
[27] C. Schmit, M. D. A. Rounsevell, and I. La Jeunesse. The limitations of spatial land use data in environmental analysis. *Environmental Science & Policy*, 9(2):174–188, April 2006. ISSN 1462-9011. doi: 10.1016/j.envsci.2005.11.006. URL `http://www.sciencedirect.com/science/article/pii/S1462901105001516`.
[28] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 34, page 35, 2000.
[29] N. Stephenne and E. F. Lambin. A dynamic simulation model of land-use changes in sudano-sahelian countries of africa (SALU). *Agriculture, Ecosystems & Environment*, 85(1–3):145–161, 2001.
[30] M. G Turner, R. H Gardner, and R. V O'Neill. *Landscape ecology in theory and practice: pattern and process*. Springer Verlag, 2001. ISBN 0387951237.
[31] P. H. Verburg, G. H. J. De Koning, K. Kok, A. Veldkamp, and J. Bouma. A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. *Ecological modelling*, 116(1):45–61, 1999.
[32] P. H. Verburg, A. Veldkamp, and L. O. Fresco. Simulation of changes in the spatial pattern of land use in china. *Applied Geography*, 19(3):211–233, 1999.
[33] P. H. Verburg, B. Eickhout, and H. Meijl. A multi-scale, multi-model approach for analyzing the future dynamics of european land use. *The Annals of Regional Science*, 42:57–77, 2008. ISSN 0570-1864. doi: 10.1007/s00168-007-0136-4. URL `http://dx.doi.org/10.1007/s00168-007-0136-4`.
[34] T. Wassenaar, P. Gerber, P. H. Verburg, M. Rosales, M. Ibrahim, and H. Steinfeld. Projecting land use changes in the neotropics: the geography of pasture expansion into forest. *Global Environmental Change*, 17(1):86–104, 2007.
[35] C A Wessman. Spatial scales and global change: Bridging the gap from plots to GCM grid cells. *Annual Review of Ecology and Systematics*, 23(1):175–200, 1992. ISSN 0066-4162. doi: 10.1146/annurev.es.23.110192.001135. URL `http://www.annualreviews.org/doi/abs/10.1146/annurev.es.23.110192.001135`.
[36] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Appendix A. Spatial cluster distribution

To get an impression of how the different aggregation methods partition the world spatially we have plotted the clusters for the case of 43 clusters (60°) derived from the original data set with grain size 0.5° (Figure A.7, A.8, A.9 and A.10). This low resolution is chosen solely to improve visualization, as for higher resolutions a clear distinction of clusters with colors would hardly be possible. Real model applications typically deal with a much higher number of spatial units.

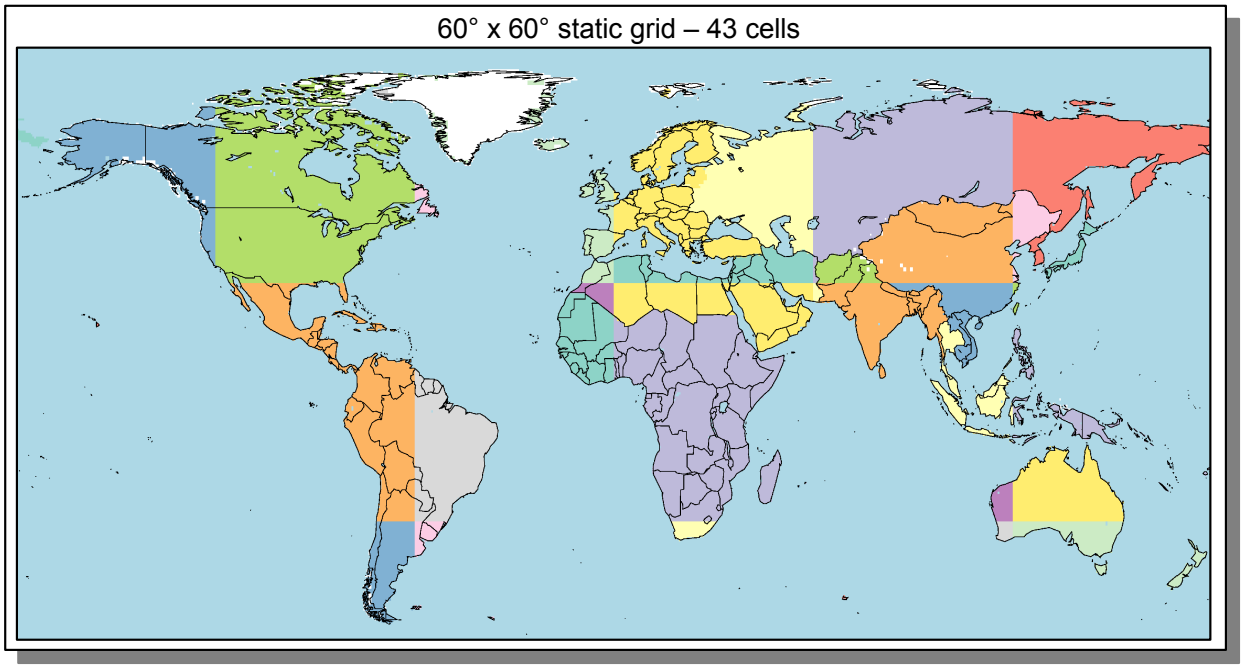

Figure A.7: Cluster distribution for regular grid aggregation

The map of data elements derived with the regular grid aggregation methods shows clearly its geometric origin (Figure A.7). The borders between the 60° squares are easy to detect. However, at many locations the geometric structure is disrupted by the country-specific world region allocation of the 10 MAgPIE regions (see model description in supplementary online material for more details). This effect is caused by the MAgPIE requirement that any cluster has to belong exactly to one world region. Therefore the 60° squares are split at region boundaries. Combined with the spatial structure of continents this leads to significant differences in cell sizes. Some data elements are huge (e.g. the green cluster in North America), other ones are tiny (e.g. the pink cluster in southern Latin America). However, this is a special case of the presented, coarse resolution. In the case of finer resolutions which are typically used for real model applications all data elements are very homogeneous in their size and distribution.

In contrast to regular grid aggregation the clusters derived with the k-means method show its strong dependence on spatial, biophysical conditions, which were used as input for the clustering (Figure A.8). Even though the location was not an explicit clustering criterion, most clusters form a big, connected cluster core with some smaller, spatially detached parts. At many locations the clusters resemble well-known geographical structures such as deserts (Australia, North Africa), but also the dependence of many biophysical characteristics on the latitude becomes visible (longish clusters parallel to the equator such as in Canada, Russia or North Africa). Clusters are distributed relatively homogeneously over the whole world.

As observed for k-means clustering, hierarchical bottom-up clustering also uncovers regions with similar biophysical conditions (Figure A.9). Some clusters are nearly identical to the k-means results (especially for Australia and Europe), others differ significantly (Russia). In contrast to k-means the clusters are distributed

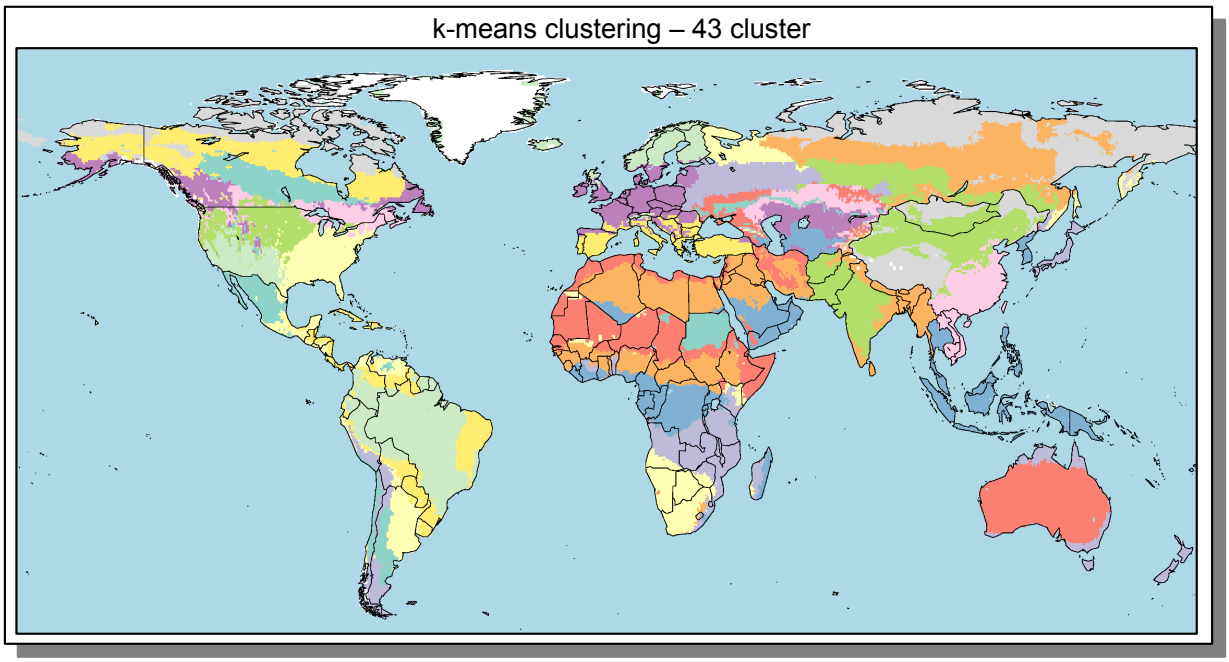

Figure A.8: Cluster distribution for k-means clustering

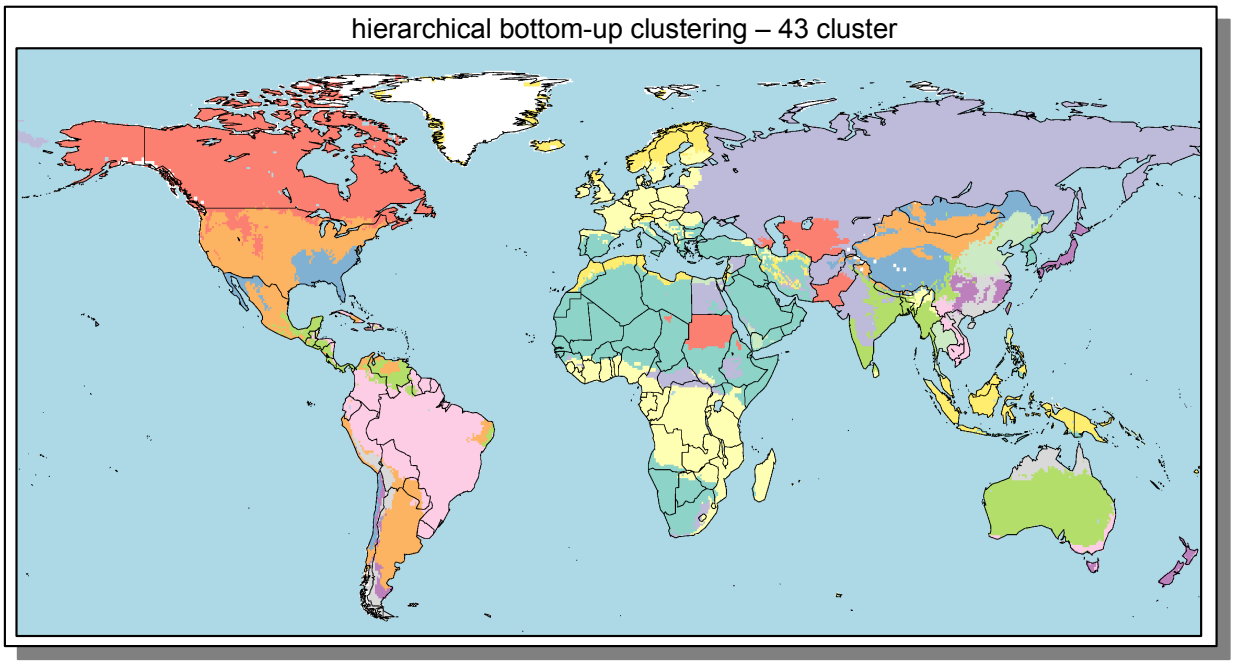

Figure A.9: Cluster distribution for hierarchical bottom-up clustering

less homogeneously over the world. Some regions, such as Pacific Asia, are clustered in more detail, other ones, such as the Former Soviet Union or North America, are represented with less clusters.

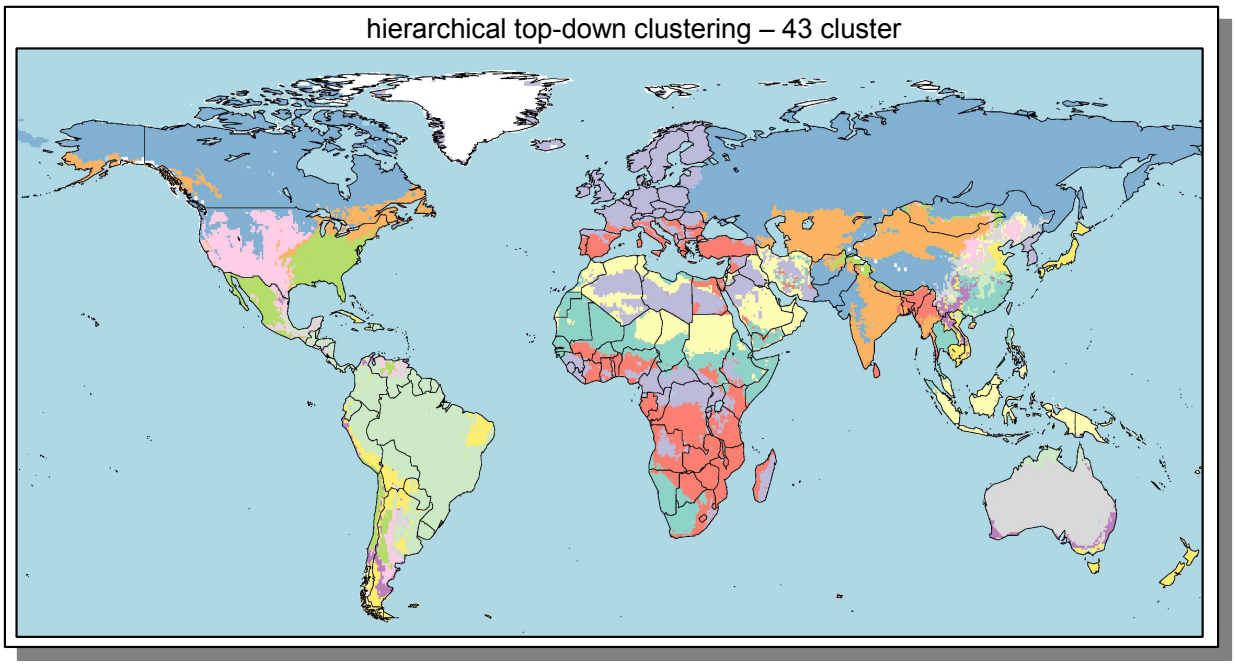

Figure A.10: Cluster distribution for hierarchical top-down clustering

The map derived with top-down hierarchical clustering (Figure A.10) shows many analogies to the result derived with bottom-up hierarchical clustering. Many structures are nearly identical, such as the huge cluster in Russia and Canada or the cluster covering Brazil. In most cases the boundaries between clusters are only slightly shifted. A systematic difference in both cluster maps is not visible.

In summary, the map illuminates the disadvantages of regular grid aggregation: geographical characteristics are not taken into account and region or continent borders lead to unwanted differences in cluster sizes. However, the latter effect occurs for very coarse grain sizes and decreases with decreasing grain size. In contrast to regular grid aggregation all the clustering methods show clear responses to the spatially explicit, biophysical conditions. Several geographical structures can be found in a similar way in all cluster maps. While both hierarchical methods deliver nearly identical results, there is a significant difference between k-means and hierarchical clustering: whereas k-means delivers a relatively homogeneous spatial distribution of clusters, hierarchical clustering shows strong differences in cluster sizes. This is most obvious for the Former Soviet Union, which is represented with only one cluster in both hierarchical maps, but with several clusters under k-means clustering. This behavior has primarily two reasons: First, the complete-linkage method which was applied for hierarchical clustering favors the creating of very small clusters, while the k-means procedure minimizing the within-sum of squares leads to more homogeneously distributed clusters. Consequently, hierarchical clustering can only apply very few clusters to the remaining areas, which have then to be combined to make very huge clusters. These huge clusters do not indicate a high homogeneity in inputs, it just means that these areas have slightly less heterogeneity then the rest of the world. Second, the number of clusters per region was predefined for k-means clustering (as the missing hierarchy does not allow for a reasonable adaption method). In contrast, the number of clusters was chosen dynamically for the hierarchical methods based on its hierarchy. K-means was forced to use more clusters for the Former Soviet Union, while hierarchical clustering was able to shift these clusters to other regions.