Ecological Modelling 263 (2013) 233–243

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

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

ARTICLE INFO

Article history:
Received 4 January 2013
Received in revised form 3 May 2013
Accepted 6 May 2013
Available online 11 June 2013

Keywords:
Aggregation
Downscaling
Clustering
Information conservation
Land use model
Scale

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.

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. (2001, 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 (Gibson et al., 2000; Turner et al., 2001; Krüger, 2007). 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 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 (Wessman, 1992; Cash and Moser, 2000; Harvey, 2000) 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 (Wessman, 1992). Second, because of nonlinearities a proper treatment of cross-scale interactions is often a requirement for accurate simulations (Cash and Moser, 2000; Harvey, 2000). 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 (Cash and Moser, 2000; Dirnböck et al., 2008).

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

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

0304-3800/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2013.05.009

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") (Lotze-Campen et al., 2008, 2009; Popp et al., 2010) 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 Sands and Leimbach, 2003, FASOM Adams et al., 1996, IMPACT Rosegrant et al., 2002), whereas models focused on geographical or ecological processes either only model certain regions of the world, with exogenous global markets (e.g. CLUE Verburg et al., 1999a,b; Wassenaar et al., 2007, SALU Stephenne and Lambin, 2001), or apply a rule-based approach (e.g. LandSHIFT Schaldach et al., 2011 – a general land use model review was done by Heistermann et al., 2006). 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., 2008). 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., 2012). 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 (Brook et al., 1988) extended with scripts for file manipulations written in PHP (Bakken et al., 2004) and scripts for calculations written in R (R Development Core Team, 2010) and Python (Van Rossum and Drake, 2000). 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) Bondeau et al., 2007), total cropland area and total land available for additional cropland expansion (Krause et al., 2009).

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

[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, etc.