Journal of Computational Science 13 (2016) 83–95

Contents lists available at ScienceDirect

Journal of Computational Science

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

Data synthesis in the Community Land Model for ecosystem simulation

Hongsheng He [a], Dali Wang [b,*], Yang Xu [c], Jindong Tan [a]

[a] Department of Mechanical, Aerospace and Biomedical Engineering, The University of Tennessee, Knoxville, TN 37996, USA
[b] Environmental Sciences Division at Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[c] Department of Geography, The University of Tennessee, Knoxville, TN 37996, USA

ARTICLE INFO

Article history:
Received 17 December 2015
Accepted 21 January 2016
Available online 10 February 2016

Keywords:
Data synthesis
Data analysis
Machine learning
Affinity Propagation
ARIMA model

ABSTRACT

Though many ecosystem states are physically observable, the number of measured variables is limited owning to the constraints of practical environments and onsite sensors. It is therefore beneficial to only measure fundamental variables that determine the behavior of the whole ecosystem, and to simulate other variables with the measured ones. This paper proposes an approach to extract fundamental variables from simulated or observed ecosystem data, and to synthesize the other variables using the fundamental variables. Because the relation of variables in the ecosystem depends on sampling time and frequencies, a region of interest (ROI) is determined using a sliding window on time series with a predefined sampling point and frequency. Within each ROI, system variables are clustered in accordance with a group of selective features by a combination of Affinity Propagation and *k*-Nearest-Neighbor. In each cluster, the unobserved variables are synthesized from selected fundamental variables using a linear fitting model with ARIMA errors. In the experiment, we studied the performance of variable clustering and data synthesis under a community-land-model based simulation platform. The performance of data synthesis is evaluated by data fitting errors in prediction and forecasting, and the change of system dynamics when synthesized data are in the loop. The experiment proves the high accuracy of the proposed approach in time-series analysis and synthesis for ecosystem simulation.

1. Introduction

Ecosystem variables play different roles in the control and representation of ecosystem states and dynamics. With a limited number of onsite sensors, ecosystem variables are commonly observed in part since many variables are unobservable or expensive to observe using onsite sensors. The problem to address is therefore the identification of significant system variables and synthesis of unobserved system variables, in order to reduce the number of onsite sensors and save the expense of practical monitoring systems. In addition, it is common practice to explore variables in ecosystem simulation for the sake of predicting climatic changes based on incomplete onsite observation. The exploration is subject to the constraints imposed by the underlying physics of geosystem variables, such that the degree of freedom in data exploration is much less than the number of variables. Data synthesis could alleviate the difficulty in data exploration while guaranteeing the physical rationality of the data.

In general, part of the ecosystem variables dominate the dynamics of the whole ecosystem, and these fundamental system variables are commonly of great interest to ecosystem scientists because of their manifest physical meanings, e.g., sun light, vegetation root growth, and ground temperature. The other variables are typically correlated to the fundamental variables in the ecosystem. Therefore, it is feasible to synthesize dependent variables with fundamental ones, so as to reduce the number of physically observed variables. Identification of fundamental variables and data synthesis are economically and operationally beneficial in the selection and placement of onsite sensors.

This paper aims to identify fundamental system variables from simulated or observed ecosystem data, and to synthesize other variables using selected fundamental system variables. The variable synthesis can avoid unnecessary observation of dependent variables and facilitate ecosystem simulation. A modular ecosystem simulation platform[1] was developed based on Community Land

* Corresponding author.
E-mail addresses: he@utk.edu (H. He), wangd@ornl.gov (D. Wang), tan@utk.edu (J. Tan).

[1] http://cem-base.ornl.gov/CLM_Web/CLM_Web.html.

http://dx.doi.org/10.1016/j.jocs.2016.01.005
1877-7503/

Models (CLM) at Oak Ridge National Laboratory, to simulate surface energy, water, carbon, and nitrogen fluxes and state variables for both vegetated and non-vegetated land surfaces [1]. The variable synthesis methods in this paper were implemented in the current simulation platform as a plugin module that simplifies and facilitates geographical studies.

The complexity of ecosystem brings many unique challenges in data analysis and synthesis. Firstly, the relation between system variables highly depends on sampling time and observation scale. In other words, the relation is a function of time, sampling frequency, and time span. Subsequently, ecosystem variables are tightly coupled such that the change of one system variable may influence a group of dependent variables. Finally, big data obtained during longtime observation render it very difficult to discover the underlying interaction between the variables.

Data synthesis is a problem that incorporates data from a variety of sources to produce new or enhanced information about a system following basic physical principles [2]. A model-based approach was proposed in [3] for the identification and prediction of phenological attributes from satellite image time series. The Nonlinear Harmonic Model was utilized to fit intra-annual response of land cover multispectral reflectances obtained from satellite image time series. The work focuses on the problem of model fitting of a given time series. A Fourier series based approach was presented in [4] to address the data missing problem using multi-temporal analysis. A functional curve, consisting of a group of Fourier series with different coefficients, are optimally fitted to yearly observed data through least square estimation (LSE). Recent work [5] presented a procedure for producing temporally smoothed and spatially complete NASA' s Moderate Resolution Imaging Spectroradiometer (MODIS) data sets. A data series was smoothed, and gaps in the series were filled to generate high-quality data from observations with missing points. From time series observed by coarse-spatial-resolution and hyper-temporal earth satellites, the land cover changes were detected automatically using different clustering methods and feature extraction processes [6]. In that paper, short term Fourier transform coefficients were computed over subsequences of MODIS data within a temporal sliding window, and meaningful sequential time series were extracted for analysis and change detection. A function fitting method was proposed in [7] to discover seasonality in time series. The method was based on nonlinear least squares fits of asymmetric Gaussian model functions directly to the time series. Data fitting methods have gained success in system data analysis and prediction [8,9]. These fitting methods, however, cannot be directly applied to ecosystem data synthesis, because many geosystem variables are physically heterogeneous, and inherent properties of geosystem variables are not directly observed in the time domain.

A similar concept that relates to this paper's work is data assimilation, which incorporates observations into a computing model of a real system. Data assimilation is used to estimate variables that are not directly observed from space but are needed for applications [10]. Data assimilation technique was utilized to estimate model parameters from time-series observations to modify the pathways while preserving model complexity [11]. The work [12] demonstrated that data assimilation combining different observations with a dynamics model improved the understanding of ecosystem carbon exchange. An ensemble Kalman filter was used to associate time series with a box model of carbon transformations. The paper [13] proposes an automatic time-series generation using ranked data quality indicators and stepwise temporal interpolation of short data gaps. Pixel-level data are employed to filter time series and interpolate invalid data with statistical or contextual methodologies.

The unique problem to solve in this paper is to synthesize unknown or unobserved yet intensely dependent variables using predefined, observed, or measured data in ecosystem simulation and prediction. This paper utilizes machine learning algorithms to better understand the behavior of the ecosystem and to bridge the gap between the geosystem simulation and onsite observation. Instead of direct synthesis of time series, the paper synthesizes data using variables with similar features that are categorized in the same cluster, to improve the fitting accuracy of models with reduced complexity.

The scheme of the proposed method is visualized in Fig. 1, which illustrates the main components of the framework: data sampling, feature extraction, data clustering, and data fitting. Interested time series are firstly resampled by a sliding window in different sampling regions and sub-sampling frequencies. Features in time and frequency domain are then extracted from the resampled time series, and configured into a hybrid feature according to geoscientists' interest. A fused clustering algorithm of Affinity Propagation and *k*-Nearest-Neighbor is utilized to classify the feature into clusters. In each cluster, a set of fundamental variables are selected to synthesize other variables. We propose to use a linear regression model with ARIMA errors to describe the relation between fundamental variables and the others to synthesize.

The main contribution of the paper is a novel framework of data analysis and synthesis, which was implemented as a module in the current CLM-based modular ecosystem simulation system. The paper proposes an algorithm to synthesize time series by clusters, where ecosystem variables with similar attributes are grouped together, instead of direct fitting in time domain. Specifically,

1. the paper proposes a feature extraction method from time series, which is customizable for different physical properties in time and frequency domain;
2. the paper proposes a data synthesis method within clusters using Affinity Propagation and linear fitting;
3. the paper recovers the physical meanings of geosystem variables in different feature space, and models the underlying relation of the variables.

2. CLM-based modular ecosystem simulation

The Community Land Model (CLM) within Community Earth System Model, developed by NSF and DOE, simulates surface energy, water, carbon, and nitrogen fluxes and state variables for both vegetated and non-vegetated land surfaces [14]. The CLM-based simulation is designed to understand the way that natural and human changes in ecosystems affect the climate. Within CLM, biogeophysical and biogeochemical processes are represented in the simulation on a hierarchical landscape surface data structure: grid cell, land unit, column, and Plant Function Type (PFT) independently. Water, energy, flux and each sub-grid unit maintain its own prognostic variables. The same atmospheric forcing is used to force all sub-grid units within a grid cell. The surface variables and fluxes required by the atmosphere are obtained by averaging the sub-grid quantities weighted by their fractional areas. The dynamics of CLM is difficult to understand because of its large amount of sub-models and global variables. The response of the CLM to a simulated environmental stimulus is unclear though the dynamics of a module is well studied. The flow of information and propagation of module-level interaction is intractable, especially in extreme conditions.

The paper focuses on the variables in the CanopyFluxes module of the developed CLM-based simulation platform. The ecosystem variables in the CanopyFluxes module is described in Table 1 with explanations of their physical meanings. According to the