A surface cover change detection method based on the Australian Geoscience Data Cube

P. Tan, S. Sagar, N. Mueller, L. Lymburner, M. Thankappan and A. Lewis

National Earth and Marine Observations, Environmental Geoscience Division, Geoscience Australia
Email: Peter.Tan@ga.gov.au

Abstract: We describe a surface cover change detection method based on the Australian Geoscience Data Cube (AGDC). The AGDC is a common analytical framework for large volumes of regularly gridded geoscientific data initially developed by Geoscience Australia (GA), the National Computational Infrastructure (NCI) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). AGDC effectively links geoscience data sets from various sources by spatial and temporal stamps associated with the data. Therefore, AGDC enables analysis of generations of consistent remote sensing time series data across Australia.

The Australian Reflectance Grid 25m is one of the remote sensing data sets in the AGDC. The data is currently hosted on the high performance computational cloud at the National Computational Infrastructure. Our change detection method takes advantage of temporally rich data in the AGDC, applying time series analysis to identify changes in surface cover.

The aim of this study is to develop a modelling framework addressing these issues, and to improve the efficiency and effectiveness of data modelling processes for the AGDC. The framework adopts a modular design, taking advantages of standardisation of data structures provided by the AGDC. The basic unit in the framework is a modelling module, which applies generic statistical functions or machine learning algorithms on a spatial-temporal partition of remote sensing data. Under the framework, a typical workflow of a modelling process consists of a sequence of connected modules. Such modular design offers both flexibility and reusability.

To detect change we apply a series of modules, which are independent of each other. The modules include:

- a pixel quality mask and time series noise detection mask, which detects and filters out noise in data;
- classification modules based on a random forests algorithm, which classifies pixels into specific objects using spectral information;
- training modules, which create classification modules using known surface cover data;
- time series analysis modules, which model and reduce time series data into coefficients relevant to change detection targets;
- temporal and spatial classification modules, which classify pixels into predefined land cover classes.

This paper summarises development of the work flow and the initial results from example applications, such as reforestation / deforestation detection and coastal zone mapping.

Keywords: *Surface cover, time series analysis, change detection*

1. INTRODUCTION

Scientists often need to extract information from satellite data when they develop remote sensing applications. The success of the applications rely on high quality remote sensing data and a modelling framework capable of discovering target information. As more and more remote sensing data become available online, time series data modelling plays an increasingly important role in tackling real world problems, such as mapping and monitoring floods and bush fires or changes in forest extent.

Remote sensing data have been playing a key role in monitoring land cover and land use changes (Brink & Eva, 2009; Dewan & Yamaguchi, 2009). An Australian example is the National Dynamic Land Cover Dataset (DLCD) (Lymburner et al. 2011) developed by Geoscience Australia (GA) and the Australian Bureau of Agriculture and Resource Economics and Sciences (ABARES). The DLCD provides a baseline for identifying and reporting on change and trends in vegetation cover and extent. With nearly 30 years of continuous Earth observation data over Australia, Landsat satellites provide essential information for mapping land cover and detecting surface cover changes.

The Australian Geoscience Data Cube (AGDC) (Purss et al. 2014) is a common analytical framework for large volumes of regularly gridded geoscientific data initially developed by GA, CSIRO and National Computational Infrastructure (NCI). The AGDC ingests geoscience data sets from various sources into gridded cells based on spatial and temporal stamps associated with the data. A spatial-temporal partition of remote sensing data can be retrieved in a multi-dimensional array format efficiently from the AGDC. Such features enable analysis of generations of consistent remote sensing time series data across Australia. The long term historical Australian Landsat data record is the first data collection to be ingested into the AGDC.

Prior to development of the AGDC, modelling for remote sensing applications usually consisted of the following steps: (a) collect satellite scenes covering target area; (b) apply spatial masks to exclude irrelevant pixels; (c) align all scenes to generate time series for each pixel; (d) calibrate spectral data, so that the values are consistent spatially and temporally; (e) conduct statistical modelling; and (e) apply modelling algorithm on the data. Such an approach may work well for small projects, however, it can be time consuming and impractical for applications covering large areas. As the volume of data grows exponentially with size of the area and the length of time, the data needs to be divided into multiple subsets, so that subsets can be handled by modelling programs. Such partitioning of data is often done in an ad-hoc and inefficient manner. As data source and target concepts change, modelling workflows and modules often need to be redeveloped. As such, algorithms and codes developed for previous applications cannot be reused in new projects.

The aim of this study is to develop a modelling framework addressing these issues, and to improve the efficiency and effectiveness of data modelling processes for the AGDC. The framework adopts a modular design, taking advantages of standardisation of data structures provided by the AGDC. The basic unit in the framework is a modelling module, which applies generic statistical functions or machine learning algorithms on a spatial-temporal partition of remote sensing data. Under the framework, a typical workflow of a modelling process consists of a sequence of connected modules. Such modular design offers both flexibility and reusability. Many modelling processes have identical parts of workflow, such as noise filtering modules and surface object classification modules. The modules can be reused in various applications as long as the underlying data format remains consistent. Flexibility of the framework first comes from modular nature of the design, which constructs a complex modelling process workflow by alternating connections of basic modules. Secondly, modules are flexible by parameterizing the underlying functions, for example, a random forest classification module can perform different classifications by plugging in different trained forests; or a time series change detection module can target different surface cover change phenomena by changing the function parameters.

The paper is structured as follows: Section 2 summarises development of the work flow, details of the modules and surface cover change detections methods using the proposed framework; Section 3 demonstrates the initial results from example applications, such as forest change detection and coastal zone mapping; Section 4 is the summary and conclusion.

2. METHODOLOGY

This Section describes workflows and components of the modelling framework. Different modules have been developed for the modelling work. These modules are independent of each other; each module completes a generic statistical or machine learning function. The modules include:

- a pixel quality mask and time series noise detection mask, which detects and filters out noise in data;

- classification modules based on a random forests algorithm, which classifies pixels into specific objects using spectral information;
- training modules, which construct classification modules by feeding known surface cover data into machine learning algorithms;
- time series analysis modules, which model and convert time series data into coefficients relevant to change detection targets;
- temporal and spatial classification modules, which classify pixels into predefined land cover classes.

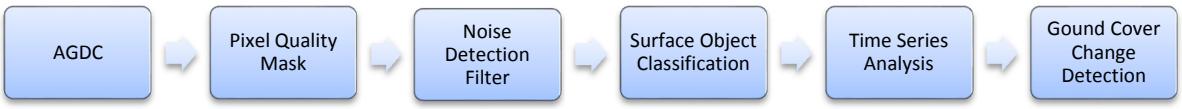

Figure 1. A typical workflow for modelling.

Figure 1 shows a typical workflow for surface cover change detection modelling. In the first step, remote sensing time series is downloaded from the AGDC. In the next step, pixel quality masks and time series noise detection filters are applied to the data to filter out noisy and invalid values among the data. The cleaned data is then fed into a surface object classification module, such as, decision tree or random forest. The next module in the workflow analyses the time series of surface objects and extracts features related to the modelling target. A surface cover change detection module is used to find location and timing of targeted surface cover change events. In the rest of the Section, details of the modules are discussed.

2.1. Australian Geoscience Data Cube

The AGDC is a common analytical framework for large volumes of regularly gridded geoscientific data. . The motivation behind the AGDC is to resolve difficulties associated with modelling large volumes of remote sensing data. Moderate resolution remote sensing data, such as that from Landsat, is often irregular spatially (the satellite pass is not parallel to either latitude or longitude) and temporally (time between two observations over a particular location is not constant). When the data is stored in conventional scene based data structure, there are significant overheads in preparing data for projects involving long time series analyses or large areas. Therefore such an approach is not suitable for large remote sensing data modelling projects.

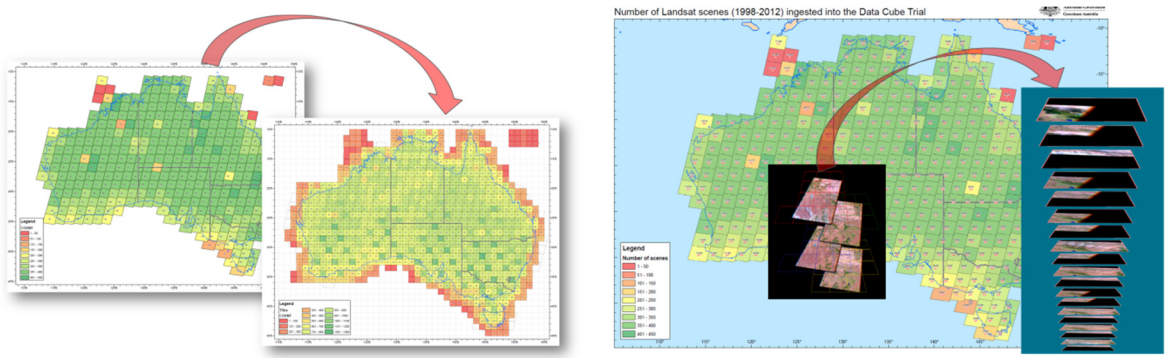

Figure 2. AGDC data ingestion.

Figure 2 shows how irregular scene based remote sensing data is ingested into the AGDC. The AGDC divides a plate into regular gridded cells, whose edges are parallel to the latitude and the longitude. Data produced by a satellite pass is ingested into cells that intersect with the pass. As such, a cell contains many tiles, each of which has a unique time stamp. In effect, a cell in the AGDC is a spatial-temporal partition of a remote sensing data set, which can be retrieved as a multi-dimensional array efficiently via an Application Programming Interface (API) provided by the AGDC.

The Australian Reflectance Grid 25m is among the first remote sensing data sets ingested into the AGDC. It includes multispectral data from Landsat 5, 7 and 8 satellites. With nearly 30 years of Earth observation data over Australia, the data is currently hosted on the high performance computational cloud at the National Computational Infrastructure. The developed change detection method takes advantage of temporally rich data in the AGDC, applying time series analysis to identify changes in surface cover.

2.2. Surface cover object classification using random forest algorithm

A surface cover object classification model is a mathematical function which takes the spectral value of remote sensing data and outputs a surface cover category label, representing a surface object class. Depending on the application, the classification outputs can be water vs non-water, green vs bare or managed land vs nature, etc. For multiclass classification, a hierarchical classification scheme consisting of a tree of binary classification modules can be used.

A classification model is usually obtained by feeding a set of training samples into a machine learning algorithm, which constructs the desired classification model. Ideally, the set of training samples and the population data should be independent and identically distributed. Random forest (Breiman 2001) is a high performance classification algorithm and has been successfully used in remote sensing classification (Pal 2005). It belongs to the family of ensemble learning algorithms. An ensemble classification model consists of a set of weak classifiers generated by a base learner. Each weak classifier may not have very high accuracy. However, ensemble algorithms achieve higher accuracy by constructing a set of classifiers which are complementary to each other. In such a set, errors made by a weak classifier are not shared by the majority of other weak classifiers in the ensemble. The final output is obtained by voting of the weak classifiers in the ensemble. If the classifier is capable of producing probabilistic output, the probability of instance below to class C given a set of data D can be calculated as $P(C|D) = \frac{1}{N}\sum_{k=0}^{N} P_k(C|D)$.

There are many algorithms capable of constructing ensemble classifiers (Dietterich 2000). Most of ensemble learning algorithms rely on injecting randomness in the base learner. In this study, the implemented random forest module create the ensemble by randomized a set of key parameters (P_a, P_b, N_c, H) in a decision tree learning algorithm (Tan and Dowe 2005), where P_a is the probability of selecting the optimal variable at an internal note, P_b is the range which the optimal cut point can be picked, N_c is the number of candidate cut points and H is the height of a tree. In this study, the random forest classification module uses 6 spectral bands (Red, Green, Blue, NIR, SWIR1 and SWIR2) and 10 derived band ratios, such as NDVI, as input.

2.3. Analysis of unevenly spaced Time series

An unevenly spaced time series is a sequence of observation value and time pairs (x_n, t_n) whose the spacing of observation times is not constant. As most common analysis methods are built for time series with constant spacing, one way to analyze unevenly spaced time series is to interpolate the data in the gaps. For continuous value time series, various algorithms can be implemented to interpolate the missing data, such as linear interpolation (Chow & Lin 1971) and Bayesian interpolation (Nieto-Baraja & Sinha 2015). For discrete value time series, nearest neighbor interpolation (Yakowitz 1987) could be implemented.

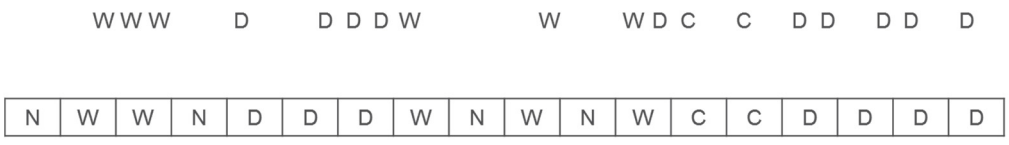

Figure 3. Interpolate unevenly spaced time series data.

Figure 3 shows the procedure of converting an unevenly spaced time series data into a regular spaced time series. On top of the figure is the time series of surface objects class $\{(c_i, t_i)\}$, where c_i is the surface cover class and t_i is the corresponding time stamp. The procedure consists of following steps;

1. Divide the timeline into regular intervals
2. Project (c_i, t_i) into the new timelines
3. Interpolate missing data by the nearest neighbour class
4. Resolve conflicts of data by a rule set, which is created by consulting remote sensing professionals with domain knowledge

These steps transform an unevenly spaced discrete time series $\{(c_i, t_i)\}$ into a regular time series $\{c_j\}$, as time stamp is no longer needed.

2.4. Time series change detection

In time series analysis, a change detection is to identify times when the probability distribution of time series changes. In remote sensing applications, one approach is to apply harmonic analysis model on time series of spectral or derivation of spectral values to detect phenological changes. BFAST (Verbesselt et al. 2010) is one example of such an approach. Harmonic analysis models work well on remote sensing time series which shows strong seasonal patterns. However, the method is not good at detecting surface cover change events which do not follow seasonal patterns, such as floods, bush fires, afforestation/deforestation, tides etc.

In this paper, we adopt another approach, which does not attempt to find the change point using spectral time series data directly. Instead, spectral data are feed into a surface cover classification model to identify surface object class. As such, time series of spectral data are converted to time series of surface object classes. The developed change detection algorithm then models the time series of objects or a set of coefficients (Tan et al. 2011) derived from the time series. Define a function $f(x,t,w)$ on a sliding window of the time series, where x is the time series data, t is the time value, w is the width of the window. Note $f(x,t,w) \in (a,b)$ as $f_{(a,b)}$, then the posterior probability of a change event happens is given by $P(C|f_{(a,b)}) = \frac{P(f_{(a,b)}|C)P(C)}{P(f_{(a,b)})}$. The prior probability of change is a constant to $f_{(a,b)}$, while $P(f_{(a,b)}|C)$ can be estimated by a set of training samples provided by remote sensing scientists, and $P(f_{(a,b)}) = \frac{\sum_a^b I(f)}{\sum_{-\infty}^{+\infty} I(f)}$.

3. APPLICATIONS

3.1. Coastal tidal zone mapping

In the inter-tidal zone, the classification model can be used to detect inundated regions, through a pixel based classification to determine the presence or absence of water. The random forest classifier is an effective technique to deal with potentially confounding scenarios in this environment, for example, partially inundated tidal flats, saturated sediments and dark, wet, exposed vegetation. The successful application of the classification model is improved with the incorporation of regionalised training data, and by a combination of reflectance signatures and indices to contribute to the random forest construction.

Through an application to the complete Landsat time series in the AGDC archive, we are able to construct a probabilistic inundation map of the intertidal zone. This not only provides a relative morphology of the observed inter-tidal region, but by attributing individual Landsat observations with knowledge of the tidal phase and offset, enables us to map the spatial extents of the inter-tidal zone observed in the AGDC archive.

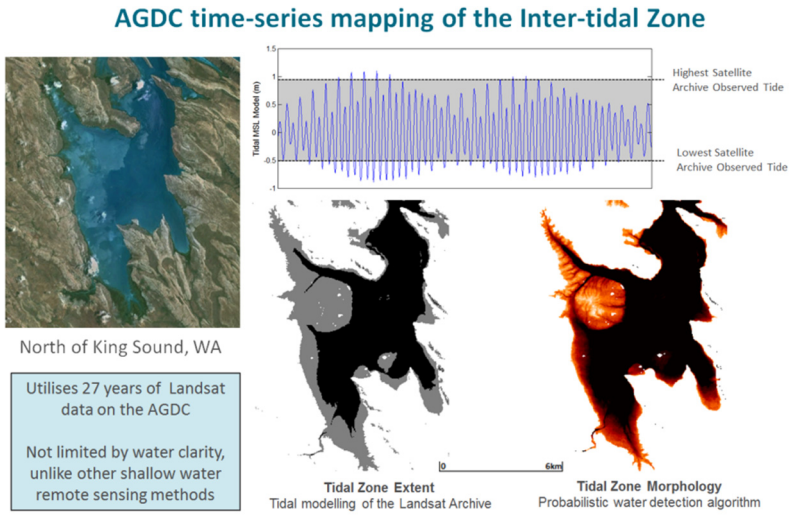

Figure 4. Mapping inter-tidal zone using AGDC data.

Figure 4 shows an example of mapping inter-tidal zone using 27 years of Landsat time series data on the AGDC. The diagram at the bottom right corner provides a relative representation of inter-tidal zone

morphology, showing areas which are exposed often in the tidal cycle (lighter colours) and areas which are only exposed at the lowest tides captured in the archive (darker brown/red shades).

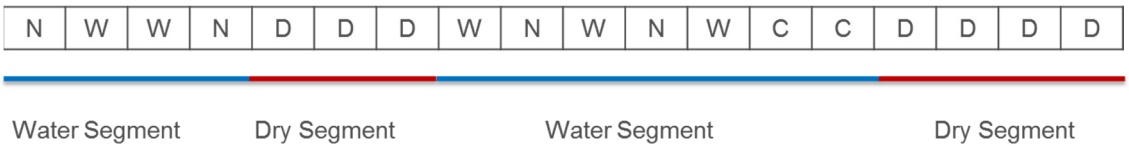

Figure 5. Calculating the relative depth of a location.

The relative depth of a location is determined by calculating the portion of time of the pixel being inundated. As shown in Figure 5, relative depth $d = \frac{\sum W_i}{\sum D_i + \sum W_i}$, where W_i is the length of a water segment and D_i is the length of a dry segment in the time series.

3.2. Forest cover change detection

A prototype change detection algorithm was developed as part of a project for the Department of the Environment (DOTE) to demonstrate the capability in three areas of interest identified by DOTE. These areas were identified as areas where forest cover change is known to have occurred and therefore provided a test framework for evaluating the performance of the change detection methodology.

The results were presented to the DOTE in early March 2015. The results demonstrated that the prototype is capable of detecting various landcover changes, including afforestation and deforestation events.

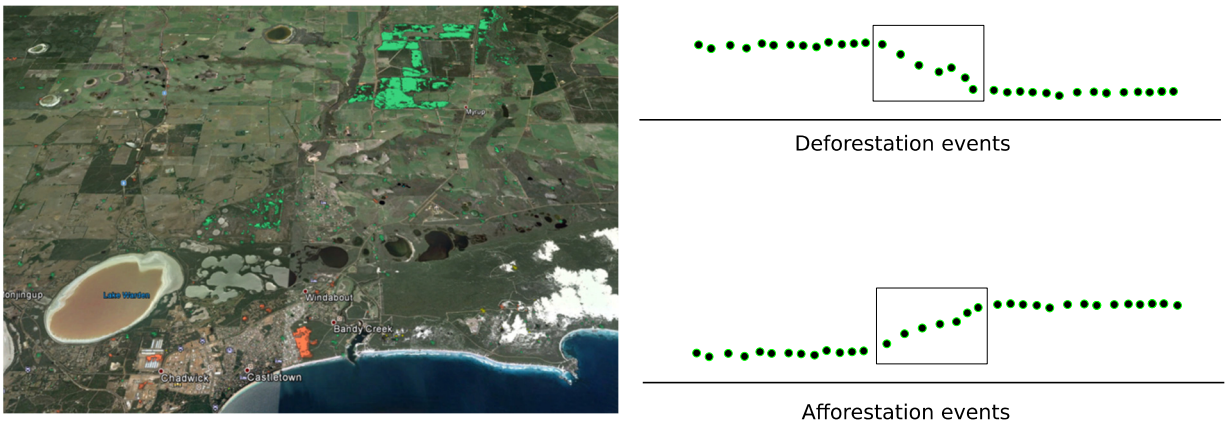

Figure 6. Time series change detection for deforestation/afforestation events.

Figure 6 shows the modelling results. The modelling algorithm successfully detects afforestation events in the target area (green pixels) and deforestation events (orange pixels). The right panel shows how a sliding window on a time series detects these events.

4. SUMMARY AND CONCLUSION

We develop a statistical modelling framework for the AGDC. The framework divides a statistical modelling process into a sequence of connected modelling modules, each of which completes a generic statistical or machine learning function. As shown in Section 3, the modelling framework has successfully supported time series analysis and change detection modelling to two remote sensing applications. The framework offers flexibility, scalability and reusability for the modelling process. Together with the AGDC, the modelling framework enables rapid development of remote sensing applications, such as landcover classification and surface cover change detection.

Recently, big data and deep learning research have been making huge progress (Najafabadi et al. 2015). The deep neural network (DNN) approach is the state-of–the-art for classification. Compared to classification

algorithms, such as random forest and support vector machine, the DNN has several advantages (Bendigo et al. 2007), such as, dealing with larger number of input variables, performing feature selection and feature creation using semi-supervised learning, and conducting multiclass classification. When computational facilities such as General-purpose computing graphics processing units (GPGPU) become available routinely, developing DNN classification and regression modules for the AGDC will further enhance the modelling capability for remote sensing applications.

ACKNOWLEDGMENTS

The author would like to thank Dr. Jin Li and Mr. Lan-Wei Wang for reviewing the paper. This paper is published with the permission of the CEO, Geoscience Australia.

REFERENCES

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*, 153.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Brink, A. B., & Eva, H. D. (2009). Monitoring 25 years of land cover change dynamics in Africa: A sample based remote sensing approach. Applied Geography, 29(4), 501-512.

Chow, G. C., & Lin, A. L. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. The review of Economics and Statistics, 372-375.

Dewan, A. M., & Yamaguchi, Y. (2009). Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. Applied Geography, 29(3), 390-401.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In multiple classifier systems (pp. 1-15). Springer Berlin Heidelberg.

Lymburner, L., Tan, P., Mueller, N., Thackway, R., Thankappan, M., Islam, A, Lewis, A., Randall, L. and Senarath, U. The National Dynamic Land Cover Dataset. Geoscience Australia, 2011.

Nieto-Barajas, L. E., & Sinha, T. (2015). Bayesian interpolation of unequally spaced time series. Stochastic Environmental Research and Risk Assessment, 29(2), 577-587.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1-21.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.

Purss, M. B., Lewis, A., Oliver, S., Ip, A., Sixsmith, J., Evans, B., & Chan, T. (2015). Unlocking the Australian Landsat archive–from dark data to high performance data infrastructures. GeoResJ, 6, 135-14

Tan, P. J., & Dowe, D. L. (2005). MML inference of oblique decision trees. In AI 2004: Advances in Artificial Intelligence (pp. 1082-1088). Springer Berlin Heidelberg.

Tan, P., Lymburner, L., Thankappan, M., & Lewis, A. (2011). Mapping cropping practices using MODIS time series: harnessing the data explosion. Journal of the Indian Society of Remote Sensing, 39(3), 365-372.

Verbesselt, J., Hyndman, R., Zeileis, A., & Culvenor, D. (2010). Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. Remote Sensing of Environment, 114(12), 2970-2980.

Yakowitz, S. (1987). Nearest-Neighbour methods for time series analysis. Journal of time series analysis, 8(2), 235-247.