ARTICLE IN PRESS
RSE-10131; No of Pages 13
Remote Sensing of Environment xxx (2017) xxx–xxx
Contents lists available at ScienceDirect
Remote Sensing of Environment
journal homepage: www.elsevier.com/locate/rse

Unsupervised detection of contextual anomaly in remotely sensed data

Qi Liu[a,*], Rudy Klucik[a], Chao Chen[b], Glenn Grant[c,d], David Gallaher[d,e], Qin Lv[b], Li Shang[a]

[a] Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, CO 80309, United States
[b] Department of Computer Science, University of Colorado, Boulder, CO 80309, United States
[c] Department of Geography, University of Colorado, Boulder, CO 80309, United States
[d] Cooperative Institute for Research in Environmental Sciences, Boulder, CO 80309, United States
[e] National Snow and Ice Data Center, Boulder, CO 80309, United States

ARTICLE INFO

Article history:
Received 2 July 2016
Received in revised form 29 November 2016
Accepted 27 January 2017
Available online xxxx

Keywords:
Unsupervised anomaly detection
Contextual anomaly
Data quality
Rare event detection

ABSTRACT

Massive amounts of remotely sensed data are being generated at an unprecedented rate, offering new opportunities for data-driven scientific discovery in the Earth sciences and related domains. However, due to the sheer volume of remotely sensed data and the lack of effective data analytics tools, most data remain in the dark, with little to no quality assurance and limited access to high-level analytical tools. Anomaly detection, which aims to find scenarios that differ from the norm, is of particular importance when analyzing remotely sensed data. However, most previous work has focused on identifying individual anomalies, and required prior knowledge of the ground truth for supervised learning. In this work, we propose an unsupervised anomaly detection framework that requires no prior knowledge and is capable of detecting anomalous events, which we define as groups of outlier objects differing contextually from their spatial and temporal neighbors. Such contextual anomalies can be useful in discovering both hidden quality issues in the data and real natural events of significance. We demonstrate the effectiveness of our framework via Web-based tools developed for visualizing and analyzing such contextual anomalies, using two types of data. The techniques and tools developed in this project are generally usable for a diverse set of satellite products and will be made publicly available with the support of the National Snow and Ice Data Center (NSIDC).

1. Introduction

Recent advances in remote sensing technology have revolutionized the way remotely sensed (RS) data is acquired, managed, and analyzed (Ma et al., 2015; Rathore et al., 2015). More than 200 on-orbit satellites are currently capturing continuous Earth observations (Ma et al., 2015), offering great opportunities for advancing the scientific understanding of the Earth's systems. However, as the proliferation of these products increases, so does the complexity needed for processing them. The variety of data can vary greatly, even within individual data sets (Li et al., 2016). Therefore, human expert-driven data analysis, a laborious and time-consuming process, remains the mainstream approach for data quality assessment (Isaac and Lynnes, 2003; Gonzalez and Datcu, 2011; Borg et al., 2011 and scientific knowledge discovery (Steffen et al., 2004; Ferguson and Villarini, 2012). The sheer volume and complexity of RS data have hampered adequate quality assessment or higher-level analysis such as anomaly detection. While Earth scientists are very interested in studying anomalies such as climate extremes (Coumou and Rahmstorf, 2012; McCright et al., 2014; Easterling et al., 2000; Muster et al., 2015), finding all such anomalies from massive data sets is challenging. Furthermore, RS data is often contaminated with noise or errors which need to be identified and then either corrected or eliminated. Thus, a high demand exists for effective and generic anomaly detection tools which require minimal involvement of domain experts while having the ability to adapt to diverse data sets. Anomaly detection in RS data is challenging for several reasons. (1) Prior models may not exist for determining what constitutes anomalous data. Additionally, unknown types of anomalies may exist in the data. (2) Remotely sensed imagery is often contaminated with noisy pixels or missing data. (3) The dynamic nature of spatial and temporal variations in multiple frequency channels need to be considered. (4) Due to the high volume and variety of

* Corresponding author.
E-mail addresses: qi.liu@colorado.edu (Q. Liu), klucik@colorado.edu (R. Klucik), chao.chen@colorado.edu (C. Chen), glenn.grant@colorado.edu (G. Grant), david.gallaher@nsidc.org (D. Gallaher), qin.lv@colorado.edu (Q. Lv), li.shang@colorado.edu (L. Shang).

http://dx.doi.org/10.1016/j.rse.2017.01.034
0034-4257/

RS data, validated ground truth data sets are not normally available for supervised learning. Additionally, there will always exist unusual anomalies in the data that exceed the expectations or prior knowledge of Earth scientists. Unsupervised approaches are thus preferred. In this work, we propose a clustering-based framework for anomaly detection, which requires no domain knowledge of the data set and enables automated anomaly detection on diverse data sets.

While most previous research has focused on detecting *point anomalies* (Chandola et al., 2009; Gupta et al., 2014; Bhaduri et al., 2011), which are individual data points that are considered globally anomalous (e.g., extreme low temperature or high wind), our work focuses on the less-studied *contextual anomalies* (Sun and Chawla, 2004; Alvera-Azcárate et al., 2012), especially in the dynamics of spatial and temporal domains. Contextual anomalies are relative anomalies under specific contexts. For example, a high air temperature trend in the summer may be normal, but if the same temperature trend occurs during a winter period it could potentially be due to data defects or anomalous atmospheric processes (Matthes et al., 2015; López-Moreno et al., 2014; Bokhorst et al., 2012). Such contextual anomalies are of particular importance in Earth sciences research. And an effective solution for detecting contextual anomalies should leverage both spatial and temporal coherence in localized regions. The assumption is that in a natural environment, pixels in close proximity share similar morphology and evolve gradually over time, while anomalous pixels would have low coherence with their neighbors in space and time.

One issue to note is the noise or errors in the data. Fig. 1 shows two snapshots of the Advanced Very High Resolution Radiometer (AVHRR) skin temperature data for the South Pole (Chuck et al., 2007). The data values fluctuate from one location to another, as well as over time at the same location. Despite this spatial-temporal dynamic, the data record is also contaminated by random noise from clouds, instrumentation, and missing data. To reduce the bias or disturbance from noisy data when searching for interesting anomalies, we have developed a noisy pixel filtering algorithm and integrated it with the anomaly detection framework.

Besides discovering individual objects ($n \times n$ pixels) that are contextual outliers relative to their spatial-temporal neighbors, it is also helpful to study these outliers collectively as *anomalous events*, which can potentially reveal unusual processes that lead to those outliers in the first place. Such underlying processes can either be systematic errors (e.g., sensor calibration error), which require intervention for quality control, or natural events (e.g., extreme weather condition), which may lead to new knowledge (Xiong et al., 2011; Song et al., 2007). With the knowledge that anomalous behaviors caused by systematic errors or rare natural events can spread to a wide range of regions and last for a long period of time, we aggregate spatial-temporal outliers into anomalous events within a global spatial-temporal context and report those events with a ranking of their importance. Combining all the points above, we have developed a novel clustering-based framework for unsupervised detection of contextual anomalies in remotely sensed data. Our main contributions are summarized as follows:

- The design of an unsupervised anomaly detection framework that (1) requires no prior knowledge of the data set, (2) identifies contextual outliers that differ from their spatial-temporal neighbors; and (3) groups contextual outliers into anomalous events to reveal possible underlying processes.
- Demonstration of the framework's effectiveness via Web-based tools we have developed, using two different types of remote sensing data: SSM/I passive microwave and skin (surface) temperatures derived from AVHRR data.
- Identification and validation of new data quality issues due to systematic or random errors as well as significant natural events.

This manuscript is organized as follows: Section 2 presents the problem formulation and key notations. Section 3 describes the anomaly detection framework as well as its usage scenarios. Section 4 presents the anomaly detection framework in detail. Section 5 reports our evaluation of the proposed framework and presents case study results. Section 6 gives an overview of related work. Finally, Section 7 concludes this work.

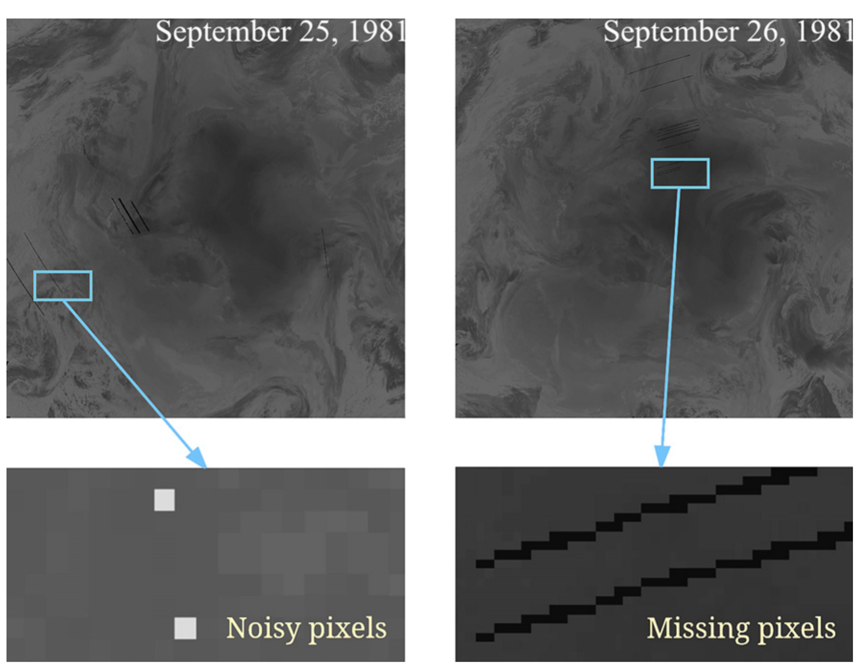

Fig. 1. AVHRR skin temperature data with noise and missing pixels. Examples shown for September 25 and 26, 1981.

Please cite this article as: Q. Liu et al., Unsupervised detection of contextual anomaly in remotely sensed data, Remote Sensing of Environment (2017), http://dx.doi.org/10.1016/j.rse.2017.01.034